James A. Sherman

Renewing Liberalism

Springer

Renewing Liberalism

James A. Sherman

Renewing Liberalism

Springer

James A. Sherman
Department of Ethics, Society, and Law
Trinity College
Toronto, Ontario, Canada

ISBN 978-3-319-28276-3 ISBN 978-3-319-28277-0 (eBook)
DOI 10.1007/978-3-319-28277-0

Library of Congress Control Number: 2016937710

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

For Lorelei
Qui novit, neque id quod sentit exprimit,
perinde est ac si nesciret.

Acknowledgements

This book began as my dissertation in the Department of Philosophy at the University of Texas at Austin. My first debt of thanks is to my supervisor, Dan Bonevac, the keenest conservative intellect I know, whose patience and encouragement enabled me to chase down my hunches and follow my arguments wherever they took me, and who was always ready and waiting with trenchant criticism when I got there. Jonathan Dancy was a sure-footed guide through some of the more rarefied terrain of ethical theory, and Al Martinich shared what was only a fraction of the impressive breadth and depth of his insight into philosophical, political, and intellectual history. Steve White ignited my interest in Aristotle's ethical and political philosophy and taught by example the virtues of rigorous scholarship. It was in his seminar on Aristotle's writings on justice, during my first year of graduate school, that the thought dawned on me that a good liberal could be a good Aristotelian. The late David Braybrooke showed me how much there was to left-wing political theory beyond the work of Rawls. Brian Leiter helped me to see the complexities involved in interpreting and defending Mill's Harm Principle, and Sahotra Sarkar helped me get my bearings in the world of decision theory. Many years ago, my fascination with social and political thought was first cultivated by John Connelly at Regis High School in New York City, and Eric MacGilvray at the University of Chicago.

I began working on this project in the winter of 2008, which I spent at the University of Oxford. Many of my views began to take shape at that time, stimulated by the wonderful discussions held at the weekly Moral Philosophy Seminar and the Jurisprudence Discussion Group. While I was there, Joseph Raz offered some illuminating commentary and well-timed encouragement for my work on moral rights and political authority. Some of the material in Chaps. 14 and 15 are taken from work done during this period and published as J Sherman (2010a) "A New Instrumental Theory of Rights" *Ethical Theory and Moral Practice* 13:2, 215–228; and J Sherman (2010b) "Unresolved Problems in the Service Conception of Authority" *Oxford Journal of Legal Studies* 30:3, 419–440.

From 2011 to 2013, I benefitted greatly from receiving a postdoctoral fellowship from the Social Sciences and Humanities Research Council of Canada, which I held at the University of Toronto. The Centre for Ethics at Trinity College, Toronto, was an ideal home during this period, and I learned a great deal from my colleagues there, particularly Joseph Heath and Thomas Hurka. The surest way to get clear on an idea is to teach it, and I have been fortunate to have had excellent students in my course on distributive justice in Trinity College's Program in Ethics, Society and Law.

My views on the wide range of topics covered in this book have been profoundly shaped by the many challenging and enriching conversations I have had with friends and colleagues in Toronto, Austin, Oxford, Chicago, and New York City. A short list, undoubtedly leaving out many, would include Reid Blackman, Francis Fallon, Joseph Forte, Conor Johnston, Maris Köpcke Tinturé, Grant Madsen, Anna-Sara Malmgren, Matt O'Brien, David Palmer, Carla Saenz, Michael Sevel, and Neil Sinhababu. Two who stand out for special mention are Blinn Combs and Eric Hochstein.

Seven years is not a short time to be engrossed in an intellectual project, and the unfailing patience and support of my wife, Jennifer Neilson, has been the *sine qua non* of its completion. The encouragement of my parents, Joan Caiazzo and Arthur Sherman, has also been indispensable. And my faithful dog Asta kept me company throughout the entire writing process. Though I have been actively at work on this book since 2008, and the seeds of the project were planted over a decade ago, my determination to write it stretches back even further. I grew up surrounded by liberals, and did not encounter a cogent and precise articulation of conservative principles until I went away to university. It was then that I realized that I did not have a good answer to the question of why I was a liberal. Whether the argument of this book provides a good reason to be a liberal is something others will judge; but it does provide the best account I can give of why I count myself as one.

Trinity College, Toronto James A. Sherman
6 January 2015

Contents

General Introduction

This book lays the foundation for a distinctively Aristotelian variety of progressive political liberalism. The theory I develop belongs to the progressive liberal tradition insofar as it endorses a progressive interpretation of the following principles:

1. *The Principle of Liberty*: The primary goal of the State is to promote, preserve, and protect individual freedom and autonomy.
2. *The Principle of Competitive Value Pluralism*: There are many equally good ways of life, which are incompatible insofar as leading one excludes leading others, and the values that structure some conflict with the values that structure others.
3. *The Harm Principle*: The only adequate justification for State interference in individuals' lives is the prevention of harm in the form of restrictions on other individuals' freedom or autonomy.

Of course, if we leave some key terms uninterpreted, these principles are shared with classical liberalism, libertarianism, and neoliberalism—views I characterize in the Historical Introduction as essentially conservative. At some level of abstraction, these principles are a common root shared by modern conservatism and modern liberalism. The views branch out according to their very different conceptions of freedom, autonomy, and the good life. As we see in the Historical Introduction, it is their conception of freedom in particular, which interprets freedom in such a way that it is not equally valuable to all individuals, that makes these modern conservative views continuous with feudal conservatism (though the latter would reject the second and third principles, and severely restrict the reference of "individual" in the first, turning it into a principle of noble privilege). My progressive interpretation of these principles is based on an Aristotelian interpretation of these key notions:

A. *Substantive Agent Freedom*: The extent of an agent's freedom is the extent of that agent's capability set—the set of ways of life he has a real opportunity to lead. These ways of life are constituted by the valuable functionings—the valuable states of being, actions and activities, and projects and goals—which the agent has a real opportunity of realizing.

B. *Agent Autonomy*: An agent is autonomous insofar as he exercises the capacity of autonomy, the core component of which is the capacity for deliberating about ends. Ideal competence in exercising this capacity is the hallmark of the Aristotelian *phronimos*.
C. *Agent Well-being*: An agent's well-being consists in his willing pursuit and achievement of valuable functionings, chosen through deliberation from an adequate range of options, within the confines of respect for his moral duties.

It is these Aristotelian ideas, which I develop and defend over the course of the book, which will guide us toward a progressive liberal political theory: one which conceives of freedom in a way that makes it equally valuable to all individuals, and one according to which equality of freedom is inconsistent with severe inequalities of wealth and socio-economic power.

The book has two divisions, each of which has two parts.

Division I: Liberty

The first division develops an account of individual liberty, a notion which I decompose into the elements of autonomy and freedom.

Part I: Liberty: Autonomy

Part I is concerned with the capacity of autonomy, the development and exercise of which are important aspects of agents' well-being. The central component of individual autonomy is the capacity to deliberate about what ends to adopt—to figure out what to value. In Chaps. 3 and 4, I develop the first ever formal, dynamic, endogenous account of rational deliberation about ends. Excellence in exercising this capacity is the hallmark of the Aristotelian *phronimos*—the fully autonomous practical agent. My account of the process an autonomous agent engages in when he determines which goals to adopt and pursue thus provides a more thorough characterization of the capacity of autonomy than has ever been offered.

In Chap. 3, I lay out the decision theoretic background against which my account is developed. I critique Henry Richardson's and Elijah Millgram's attempts to develop an account of deliberating about ends. I then turn to decision theory, and in particular the possibility of rigorous evidential and causal decision theories based on the work of Frank Ramsey, as extended by Richard Bradley and John Howard Sobel. This Ramsey-style decision theory serves as the background for the account of rational deliberation about ends I develop in the following chapter, and I defend that choice and explore the potential for a Ramsey-style theory to provide an

idealized account of instrumental deliberation (which is also an essential component of autonomy). I then reflect on some promising ideas that have been offered by Amartya Sen and others, regarding the way in which a framework for representing deliberation about ends should be constructed. I begin Chap. 4 with a discussion of the concept of an end, the relationships between ends and means of different types, and the place of ends in practical reasoning, which draws extensively from Aristotle. Finally, I develop my framework for representing ends-deliberation in some detail, drawing on Brian Skyrms' work on the dynamics of instrumental reasoning. I conclude the chapter by relating this discussion to another aspect of autonomy: authenticity.

Part II Liberty: Freedom

Part II develops an account of individual freedom in light of the theory of autonomy articulated in Part I. Over the course of Chaps. 6 and 7, I construct an enhanced version of a neo-Aristotelian capabilities-based characterization of individual freedom, building primarily on the work of Amartya Sen. I argue that we should understand individual freedom in terms of the opportunities that an agent has, first, to develop his capacity of autonomy, and then to autonomously choose to develop and exercise his capabilities over the course of his life. I show that my view of freedom overcomes a number of deficiencies and limitations in Sen's account. I then turn to the question of how individual freedom should be measured. Drawing on recent developments in social choice theory, I propose a comprehensive way of assessing how much freedom an individual has, taking both the number and the diversity of his opportunities into account. I also discuss another important dimension of personal autonomy: self-control. I put the model of ends-deliberation already developed to use in an original account of failures of self-control, a phenomenon known in the philosophical tradition as weakness of will. Finally, I argue that we should understand individual liberty, the guiding value of political liberalism, as a compound of individual autonomy and freedom, and provide a critical discussion of the politically conservative conception of freedom.

Division II: Justice

The second division is concerned with the issue of social justice from two perspectives: the question of what the most morally defensible distributive scheme is; and the question of how to both justify and limit the State's authority to enact and enforce policies aimed at realizing that scheme.

Part III: Justice: Distribution

Part III concerns the questions of what the redistributive goals of the State should be, and how those goals are to be justified. In Chaps. 9, 10 and 11, I use my compound characterization of individual liberty to argue that the equality of liberty is the appropriate goal of the State's redistributive efforts.

In Chap. 9, I assess a range of alternative approaches to distributive justice, both egalitarian and non-egalitarian. The discussion of egalitarian theories has a dual focus: what form should the theory take (strict equality, satisficing, maximizing the position of the worst-off, etc.), and what should the object of distributive concern be. With respect to the first issue, I argue that views advocating equality-of-welfare cannot avoid decisive objections stemming from the moral importance of personal responsibility. Instead, the distributive focus should be on individuals' shares of liberty in the precise sense in which this notion has been defined over the previous chapters, and that an egalitarianism of well-being subject to desert collapses into this view. I give special attention to the equality-of-resources approach to distributive justice advocated by Ronald Dworkin. I show this theory to be an incomplete version of a liberty-based egalitarian theory, not a rival to one.

In Chap. 10, I discuss the economic theory which serves as a background to my theory of social justice. I explain my reasons for rejecting neoclassical economic theory and its descendants, in an effort to immunize my theory of justice and the policies it requires from objections, which are sure to come from the neoclassical camp, that it is inconsistent with a robust and smoothly functioning national economy. I then describe the essential characteristics of the school of economic theory to which I adhere, the Evolutionary-Institutional school, and give my reasons for doing so.

In Chap. 11, which is the heart of the book, I introduce my own theory of social justice, which I call the theory of Equal Liberty. I argue that the appropriate distributive goal of the State is to equalize each individual's share of liberty, as this notion has been defined over the course of the first seven chapters. I argue that this theory of social justice satisfies a number of desiderata, including a commitment to equality, an appropriate respect for autonomous effort, and a commitment to preventing exploitation. I defend this view from a number of possible objections. After considering the issue of social justice from the perspective of a single nation and a single generation, I introduce international and intergenerational perspectives and explore the possibilities for reconciling these viewpoints. Finally, I consider the practical question of what policy implications my theory of social justice has. I argue that the policy program of the Social Market Economy, developed in the work of the German Catholic economist, social theorist, and policy-maker Alfred Müller-Armack, provides an excellent practical counterpart to the theory of Equal Liberty. I argue that pursuing the goal of Equal Liberty through this policy program is consistent with a number of other significant social goals, including economic efficiency, environmentally sustainable growth, and the protection of democracy from encroachments by economic power.

Part IV: Justice: Authority

Part IV is concerned with the basis and limits of the State's authority to pursue the sort of approach to distributive justice that I have been defending. In Chaps. 13 and 14, I develop a teleological account of individual rights, taking Joseph Raz's influential but problematic view as my point of departure. Chapter 13 focuses on the nature of moral duty and is motivated by the challenge issued to all secular moral theories by G.E.M. Anscombe over 50 years ago: to find a way to justify the use of legalistic concepts like duties, in the absence of a moral legislator. In Chap. 14, I argue that an individual's moral rights emerge from the duties that are owed to him in virtue of his interests. I discuss four features an individual's interest must have in order to ground a duty that can be justifiably enforced. The individual has a moral right in virtue of being owed such a duty. I also discuss the place of the virtues within the theory of moral duties and rights, and the compatibility between the theory of moral duty and Jonathan Dancy's moral particularism. This discussion provides an opportunity to complete the account of personal autonomy by describing a way to formalize Dancy's theory of ethical deliberation—the last type of rational deliberation crucial to autonomy. This account of ethical deliberation is then integrated with the accounts of instrumental and ends-deliberation to provide a unified view of the rational dimension of autonomy. I conclude the chapter with a defense of the theory's various commitments against the view that morality is an evolved system of social rules.

In Chap. 15, I use this account as the foundation for an account of legitimate political authority. I discuss the conditions under which a political authority's interest in maintaining social order is sufficient to ground a duty of compliance on the part of those subject to the authority. When this duty can be justifiably enforced, the authority has a right to that compliance, and is therefore legitimate. The view developed is based on Joseph Raz's widely influential theory of authority, and in particular on his Normal Justification Thesis. I defend Raz's own view against some powerful objections to it that have been raised by William Edmundson. I raise a number of major problems insolvable by Raz's theory, however, and show that my theory is able to diffuse all of them.

In Chap. 16, I address the issue of the appropriate limits on a legitimate authority's use of coercive power. The discussion is guided by a commitment to the Harm Principle—part of the bedrock of political liberalism—and begins with the issue of how this principle should be interpreted. I then scrutinize a number of objections to the compatibility between Jospeh Raz's liberal perfectionism and the principle, and argue that nearly all of these fail. One objection, however, does manage to show that Raz's own interpretation of the principle fails to justify a sufficiently narrow understanding of harm, and thus to set plausible limits on State authority. In the remainder of the chapter, I develop and defend an interpretation of the principle as prohibiting actions which threaten others' possession of an equal share of liberty. I discuss the nature of the value of liberty and argue that this understanding of harm sets appropriate limits on State action while still enabling the State to pursue the goals of Equal Liberty within the bounds of its legitimate political authority.

Chapter 1
Historical Introduction

The difficulty lies, not in the new ideas, but in escaping from the old ones, which ramify, for those brought up as most of us have been, into every corner of our minds. – John Maynard Keynes, The General Theory of Employment, Interest, and Money

This book is an attempt to escape from an old idea. Its aim is to achieve what J. K. Galbraith called "the emancipation of belief" (Galbraith 1973, p. 241). Galbraith saw the modern resident of a capitalist democracy as in need of emancipation from a particular belief. To understand the content of this belief, we must familiarize ourselves with the basic terms of his institutional analysis of modern capitalism. He used the term "the planning system" to refer to the small group of the very largest transnational corporations which have come to exercise a great deal of economic, social, and political power in the modern world; and the term "the technostructure" to refer to its senior management. The belief in question, then, is that "the purposes of the planning system are those of the individual…any public or private action that serves its purposes serves also the purposes of the public at large" (Galbraith 1973, p. 241). There is much truth in Galbraith's analysis of modern economic, social, and political life. We shall return to it at several points in the pages that follow. But his historical perspective does not extend far enough. The cultural and intellectual forces which he identifies as characteristic of the age of the post-WWII "new industrial state" emerged more than a century earlier. Indeed, they may be accurately viewed as the values of the "old industrial state," since their ascendency dates back to the age of the Industrial Revolution.

Take, for example, one of these values which Galbraith terms "the convenient social virtue":

> The virtue in question is that which is convenient to the purposes of the planning system. The virtuous head of a family works hard for an income that, however, is never quite sufficient for the things the family needs. These, as a practical matter, always increase a little more than income…If the individual is a professional or an executive, the foregoing compulsions are much increased…he is peculiarly restless in his efforts. He, of all people, cannot be negligent in his commitment to what is always called a better standard of living…
>
> The need, very simply, is recognition that our beliefs and the convenient social virtue are derived not from ourselves but from the planning system (Galbraith 1973, pp. 241–243).

Endorsement of the "convenient social virtue," however, long pre-dates the emergence of the planning system. This way of thinking arose from the conditions

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_1

of the Industrial Revolution, in the fiercely competitive *laissez-faire* environment of early-nineteenth century England. It is characteristic not only of our own age, but of the Victorian age, from which we have inherited it. It was in the Victorian era that "the creed of success, like its practitioners, rose to the top," a creed which urged each individual to concentrate his efforts on "win[ing] the race of life…reach[ing] the top and hold[ing] a position in which *you* gave the orders that others executed—this was the crowning glory" (Houghton 1957, pp. 194, 191). And such striving was for the Victorians, even more explicitly and self-consciously than for ourselves, a personal and social virtue: "It was the bounden duty of each citizen to better his social status; to ignore those beneath him, and to aim steadily at the top rung…Only by this persistent pursuit by each individual of his own and his family's interest would the highest general level of civilization be attained…" (Webb cited in Houghton 1957, p. 188). Galbraith observes that, in the modern age, "It is possible to imagine a family…which makes a considered and deliberate choice between leisure and idleness and consumption…But such esteem as it enjoys is the result, primarily, of its eccentricity." In these remarks, he unknowingly echoes the observations of the Victorian poet Robert Southey: "There may be here and there an individual, who does not spend his heart in laboring for riches; but there is nothing approaching to a class of persons actuated by any other desire" (Southey cited in Houghton 1957, pp. 183–184).

Galbraith also attributes to the influence of the planning system the modern view that the purpose of education ought to be enabling the young to become "high achievers…a close synonym for those seeking high income and consumption." He observes that this view of education picks out, as the appropriate fields of study, "engineering, science, business administration or other of the useful arts…" (Galbraith 1973, p. 242). He attributes this consequence to the close relationship between these fields of study and the needs of the planning system:

> No one will be in doubt as to the source of these attitudes. It lies with the technostructure and the planning system and with their ability to impose their values on society and the state. The technostructure embraces and uses the engineer and the scientist; it cannot embrace the artist…From these attitudes come those of the community and the government. Engineering and science are socially necessary; art is a luxury (Galbraith 1973, pp. 82–82).

But here too, the values and attitudes in question date back to Victorian culture and the influence of the emergence of industrial capitalism. J.S. Mill lamented that "philosophy—not any school of philosophy, but philosophy altogether—speculation of any comprehensive kind, and upon any deep or extensive subject—has been falling more and more into distastefulness and disrepute among the educated classes of England" (Mill cited in Houghton 1957, p. 112). As the historian Walter Houghton has documented, the attitude Mill describes was vigorously advocated by his contemporaries Charles Kingsley and James Froude:

> Kingsley thought it quite right not only that practical considerations should determine the value of any study, but that the value should be measured in pounds, shillings, and pence: "What money will it earn for a man in life?—is a question…which it is folly to despise"…

> When [Froude] succeeded Mill as Chancellor of St. Andrew's, his inaugural address was plainly a reply to his predecessor's plea for intellectual and aesthetic culture. He deplored the devotion of so much time and effort in university education "to subjects which have no practical bearing upon life…History, poetry, logic, moral philosophy classical literature, are excellent as ornament…but they will not help you stand on your feet and walk alone…the only reasonable guide to choice in such matters [of study] is utility" (Kingsley cited in Houghton 1957, pp. 119–121).

Here, Froude's conception of utility is exclusively "tangible results—profits, larger plants or firms, personal advancement, professional and social…utility in the narrow sense" (Houghton 1957, p. 111).

The purpose of this brief foray into intellectual history is to discern the fact that the beliefs, values and attitudes which Galbraith identifies, which he rightly recognizes as ones from which we are in need of emancipation, are part of a cultural inheritance stretching back to the dawn of industrial capitalism. To emancipate ourselves from them, we must first emancipate ourselves from a much deeper belief, one born in the same era, and still exercising a profound influence on Western culture today.

It is hardly original to observe that four views form the basis of modern conservative thought, as first developed in the works of Edmund Burke at the end of the eighteenth century. These are: a preference for liberty over equality; an attitude of suspicion toward the power of the State; elitism; and respect for tradition and established institutions, coupled with a cautious skepticism of the idea of progress. This first view, or more precisely its implication that there is an essential tension between the preservation of individual liberty, and the achievement of equality with respect to wealth, social status, or personal achievement, evolved, in the first half of the nineteenth century, into the belief which concerns us. This is the belief that *an extensive and equal degree of liberty is possible for every individual*, *consistent with the existence of great inequalities of wealth and social and economic power*; *and that public action which diminishes these inequalities invariably encroaches on individual liberty*. This is the core thesis of modern conservative thought, the conservative thought of the nineteenth century which has been inherited by the twentieth century and the twenty-first century.

In nineteenth century England, modern conservatism replaced the old, feudal-aristocratic conservatism of the eighteenth century and earlier as the dominant ideology.[1] Naturally enough, it was not referred to as "conservatism" during this time, but was in fact, owing to its explicit preoccupation with the idea and the value of liberty, referred to as "liberalism." This is why there are a number of contemporary views with names like "classical liberalism," "neoliberalism," and "libertarianism," which in fact belong to the modern conservative tradition. The core thesis of modern

[1] Of course, the historical details are considerably more complex than this simple formula makes them out to be. Although the modern conservative worldview achieves dominance in English culture in general during the Victorian era, it is already fully formed in the Late Puritanism of the late seventeenth and early eighteenth centuries, and it (or something much like it) can be found expressed by members of the newly emergent English middle class as far back as the early sixteenth century (Tawney 1926, ch. 3–4).

conservatism depends on a particular understanding of liberty, one which was first embraced during the Victorian age: "In the new liberal theory [i.e. in modern conservative thought] all men were free, politically and economically, owing no one any service beyond the fulfillment of legal contracts; and society was simply a collection of individuals, each motivated—naturally and rightly—by self-interest" (Houghton 1957, p. 77). As Houghton observes, "The Victorian hymn to liberty, political and economic, was distinctly addressed to middle-class liberty" (Houghton 1957, p. 46).

It may be possible, at least in principle, for there to be a society in which every individual is equally free to make any offer of contract to any other, and equally free to either accept or refuse any such offer, acquiring legal obligations only if he accepts, and only those which are explicitly stated.[2] The crucial point, however, is that this is not the only form individual freedom can take, and it is certainly not a form of freedom which is equally valuable to all individuals. It is, rather, the form which is most valuable to certain socio-economic groups, *given* the social, economic, and political organization of the society in which they find themselves. The desire to restrict the cultural understanding of liberty, and the use of the concept of liberty in social discourse, to this conception of liberty, is what places Victorian "liberalism," and the contemporary view mentioned above, in the same overarching conservative tradition as the old feudal-aristocratic conservatism. As the historian Corey Robin explains, "Though it is often claimed that the left stands for equality while the right stands for freedom, this notion misstates the actual disagreement between right and left. Historically, the conservative has favored liberty for the higher orders and constraint for the lower orders. What the conservative sees and dislikes in equality, in other words, is not a threat to freedom but its extension" (Robin 2011, p. 8). The old conservatism was open about the fact that the liberties it prized—the liberties of the aristocracy, more often referred to by the revealing term "privileges"—were by nature exclusive. Modern conservatism prizes liberties which may be possessed universally, but which overwhelmingly serve the interests of the business community.

To say that the modern conservative view of individual liberty is one which we must emancipate ourselves from is to acknowledge that efforts to ingrain it in Western culture have been successful. How do we account for this, given that it is a view which serves the interests of so few? The story of the origins of that success also belongs to the first-half of the nineteenth century; but it plays out in the United States, not in England. What was identified as "liberal" in Victorian England—the promotion of an economic, social and political environment favorable to the business class—was recognized as the very heart of conservatism in the U.S., a nation which lacked a tradition of aristocratic privilege to be overthrown by liberalizing reforms. From its founding, American conservatism was represented by the Federalist Party. The Federalists shared a deep fear with their "liberal" Victorian

[2] This was not true, however, of Victorian society, and has likely never been true of any actual society (Robin 2011, pp. 4–5).

counterparts: a fear of democracy, in the (very limited) sense of universal white male suffrage. The leading intellectual light of nineteenth century Federalism, Daniel Webster, argued in 1820 that "There is not a more dangerous experiment than to place property in the hands of one class, and political power in those of another…If property cannot retain the political power, the political power will draw after it the property" (Webster cited in Schlesinger 1953, p. 269). In England, Lord Macauley expressed precisely the same sentiment in 1842, arguing that to extend the vote to those without property was "incompatible with property and…consequently, incompatible with civilization," and that the first act of a democracy "will be to plunder every man in the kingdom who has a good coat on his back and a good roof over his head" (Macauley cited in Houghton 1957, p. 55). With the election of Andrew Jackson in 1829, however, America entered the age of Jacksonian democracy, which "destroyed neo-Federalism as a public social philosophy and restated fundamentally the presuppositions of American political life" (Schlesinger 1953, p. 267). Universal male suffrage became a reality in almost every state. The Federalist Party was defunct, and the Whig Party took up the cause of conservatism. It is this period which provided the impetus for a transformation, not in the substance of modern conservative thought, but in the public arguments that were offered for it. These arguments would now have to be addressed to, and to persuade, those whose interests are not best served by the conservative conception of liberty. The "ideals of Jackson" would have to be "reconcile[d] with the continued rule of the business classes" (Schlesinger 1953, p. 268).

This reconciliation was accomplished by means of a double strategy, which "set the case of the business community on fresh and unspoiled grounds" (Schlesinger 1953, p. 283). The defensive strategy was to replace "the class-conflict doctrines of Federalism" with "various theories of the identity of class interests" (Schlesinger 1953, p. 270). The historian Arthur Schlesinger, Jr. provides us with a glimpse of the efforts to promote these theories:

> "Never was an error more pernicious," exclaimed Dr. Robert Hare, an eminent Philadelphia scientist, "than that of supposing that any separation could be practicable between the interests of the rich and the working classes. However selfish may be the dispositions of the wealthy, they cannot benefit themselves without serving the laborer."…
>
> Not only were the interests of the classes identical, but there were, come to think of it, no classes at all in America. Daniel Webster became the chief champion of this view, as he had been the champion of the opposing view two decades before…
>
> If there were no class distinctions then there were two possibilities: everyone might be a workingman, or everyone might be a capitalist. The conservatives adopted both theories…But the second view was in the long run more popular…the paths of wealth are open to all…"Every American laborer," wrote Calvin Colton, "can stand up proudly and say, I AM THE AMERICAN CAPITALIST, which is not a metaphor but literal truth." And the conclusion? "The blow aimed at the moneyed capitalist strikes over on the head of the laborer, and is sure to hurt the latter more than the former" (Schlesinger 1953, pp. 270–271).

The offensive strategy was to craft an argument for small government—which is to say, a government that will not interfere with the sort of liberty prized by conservatism—which appealed to populist impulses:

> The issue, they said, was not class tyranny but executive tyranny. The basic conflict was not between exploiters and exploited, but between the governors and the governed. The main threat to liberty came, not from a propertied class, but from a bureaucratic class. The people should rise and rebuke the pretensions, not of wealth, but of government (Schlesinger 1953, pp. 275–276).

With this argument, American conservatism departs from the political course it had been on since the early days of Federalism: "the traditional conservative position had been to distrust the legislative and aggrandize the executive" (Schlesinger 1953, p. 277). This departure, however, did not result from any substantive philosophical change. It was a response to "the actualities of the eighteen-thirties" during which "it looked very much as if the greatest potentialities for democratic action might continue to reside in the executive" (Schlesinger 1953, p. 277). Out of political expedience "conservatism…emerged as the champion of congressional prerogative—a role it has continued to play ever since" (Schlesinger 1953, p. 277). Schlesinger goes on to note that by employing this double strategy, conservative politics entered an era of

> subterfuges and sentimentalities…Federalism and Whiggery represented the same interests in society, the same aspirations for power, the same essential economic policies; but Federalism spoke of these interests, aspirations and policies in a tone of candor, Whiggery, of evasion…its object was to promote confusion rather than comprehension. Both intended to serve the business classes, but the revolution in political values forced the Whigs to talk as if they intended primarily to serve the common man (Schlesinger 1953, p. 279).

Here, in the crisis that faced conservative thought in the nineteenth century and the way that crisis was handled, we have the ancestor of the belief identified by Galbraith, and the ancestors of the arguments still used to support that belief in our own day despite the fact that the U.S. (along with the U.K.) currently lags behind the rest of the advanced Western nations in social mobility (Blanden et al. 2005). By uncovering its historical lineage, we are able to see that acceptance of this belief depends on acceptance of an even more fundamental one: The interests of the capitalist are no different from the interests of the laborer, the interests of CEOs are no different from the interests of the working poor, *because* all are interested first and foremost in the preservation of their liberty; and the type of liberty most valued by the one group is just as valuable to the other, and can be possessed in equal measure by every member of both. Its defense should be preferred to the pursuit of any other goal; for nothing else serves so well the interests of any individual, whatever his position. As it has turned out to be a type of liberty which is consistent with great inequalities of wealth and social and economic power, these are conditions we must accept.

This book is, as I said, an attempt to emancipate us from a belief. We now have that belief before us. Emancipating ourselves from it is of immense importance at this juncture in history, at the end of a generation which has seen the waning, on a

scale which is unprecedented in modern history, of political resistance to (an equally unprecedented level of) organized wealth and socio-economic power.[3] In place of the modern conservative conception of liberty, the conception most rigorously and uncompromisingly defended in the present day by those who call themselves libertarians, I offer another. It is a conception of liberty which can be possessed equally by all. But it is a conception of liberty which is also of great value to all, not only to some. And its equal possession is not consistent with great inequalities of wealth and social and economic power. Rather, it is a conception of liberty which "posit[s] a nexus between freedom and equality" wherein "freedom and equality [are] the irreducible yet mutually reinforcing parts of a single whole" (Robin 2011, p. 9). The goal of social justice is equality of liberty, not of wealth. But achieving equality of liberty, of the sort I advocate, entails the use of redistributive policies, and the existence of a strong State that both shapes and, when necessary, participates in the market. These policies do not entail "a sacrifice of freedom for the sake of equality, but an extension of freedom from the few to the many" (Robin 2011, p. 9).

I therefore offer a theory of social justice which rejects the notion of a deep tension between safeguarding equality of liberty and severely curtailing inequalities of wealth and social and economic power. A renewed liberalism is not a liberalism that takes the side opposite to Burkean conservatism on this point, but one which refuses to sit at the table it has set. With respect to the other three basic commitments of modern conservatism, each is taken as an extreme to be avoided, just as its opposite is. In place of elitism, my theory will endorse an objective but vigorously pluralistic view of what it means to lead a worthwhile life. The authority of the strong State required by my theory is strictly circumscribed by what is necessary to achieve equality of liberty; its powers may extend no further. I temper the Burkean respect for established institutions with an insistence on learning what we can about what our institutions are and are not capable of, and willingness to act on what knowledge we acquire to improve them. I recognize the importance of tradition insofar as I do not set total independence from the values of one's community and culture as a condition of individual liberty. Liberty requires that these not be followed blindly, that they be scrutinized with a critical eye. Finally, I offer no grand utopian plans for the reformation of society from the top down—though we should remember that such schemes are not exclusive to the political left.[4] I offer a theory of social justice for the world we live in, one which recognizes that our aspirations must be realistic, and that the sorts of policies suited to preserving those aspirations once achieved may differ significantly from those needed to achieve them in the first place, given where we must start from.

[3] This state of affairs is the result of a widespread, well organized, concerted effort on the part of the partisans of highly concentrated wealth and power to win the hearts, minds, and votes of the middle class, an effort carried out under the aegis of a "free market" ideology which began after World War II and culminated in the late 1970s and early 1980s. See (Frazer 2015), especially Part II, for a superb and incredibly timely historical study. The definitive work on the incredible growth in wealth concentration and inequality in advanced Western nations over the past four decades is of course Tomas Piketty's recently published *magnum opus* (Piketty 2014).

[4] For a fascinating discussion of this political impulse, see (Scott 1999).

One methodological note before we proceed. At least some readers will notice that in many—but not all—cases, after describing the view of an author with whom I disagree, I simply state my objection to that view and move on, without engaging with the other author's position in an extended and dialectical fashion. I do this because I am engaged in a primarily *constructive*, rather then dialectical, project; and as such, the criterion I use to guide my decisions about the way to handle disagreements with my interlocutors is one which is appropriate to a project of this type. In each case, the relevant question is whether a dialectical engagement would lead to a discovery which makes a positive contribution to the main constructive project. In those cases where it will, but only in those cases, I engage with the work of the relevant author dialectically and at length.

References

Blanden, J., P. Gregg, and S. Machin. 2005. *Intergenerational mobility in Europe and North America*. Centre for Economic Performance Report.

Frazer, S. 2015. *The age of acquiescence: The life and death of American resistance to organized wealth and power*. New York: Hachette.

Galbraith, J.K. 1973. *Economics and the public purpose*. New York: Pelican.

Houghton, W.E. 1957. *The Victorian frame of mind*. New Haven: Yale University Press.

Piketty, T. 2014. *Capital in the 21st century*. Cambridge: Belknap Press.

Robin, C. 2011. *The reactionary mind: Conservatism from Edmund Burke to Sarah Palin*. Oxford: Oxford University Press.

Schlesinger Jr., A. 1953. *The age of Jackson*. Boston: Little, Brown.

Scott, J.C. 1999. *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven: Yale University Press.

Tawney, R.H. 1926. *Religion and the rise of capitalism*. New York: Harcourt, Brace & Company.

Part I
Liberty: Autonomy

Chapter 2
Liberty: Autonomy – Introduction

One plausible way to characterize autonomous agents is to claim that they reflectively endorse their preferences (Dworkin 1988; Frankfurt 1988). But how exactly should we unpack the notion of reflective endorsement? One popular suggestion is Harry Frankfurt's: an agent reflectively endorses a first-order preference when he has a second-order preference that he have that first-order preference (Frankfurt 1988). This suggestion has been met with the objection that one's second-order preferences do not necessarily reflect an authentic self any more so than one's first-order preferences do (Thalberg 1989). In response to this objection, some authors have developed limitations on the history of, and influences on, the formation of first-order preferences that only autonomous agents satisfy (Christman 1991; Mele 1995). But no one has yet succeeded in articulating a precise account of the process through which an autonomous agent determines what preferences to have, and whether to endorse them. In the chapters that follow, I will develop a more robust and more precise characterization of the capacity of autonomy—the capacity to lead an autonomous life, one that is self-ruled or self-governed—than has yet been offered. In these first three chapters, I will be concerned exclusively with the rational dimension of this capacity: the capacity to deliberate over and choose one's own ends, to form intentions to pursue those ends, and to deliberate over and choose means to achieve those ends. My particular concern will be with the first of these—rational deliberation about ends. I address the other dimension of autonomy, self-control, in Chap. 7.[1] I begin here with a brief review of the work of Gerald Dworkin and Joseph Raz on the nature and value of autonomy. An examination of their work provides an excellent starting point for my task, while also bringing the deficiencies of prior accounts of autonomy into focus.

Raz takes autonomy to be "an ideal of self-creation…An autonomous person's well-being consists in the successful pursuits of self-chosen goals and relationships" (Raz 1986, p. 370). In order to possess the capacity of autonomy, a person "must have the mental abilities to form intentions of a sufficiently complex kind,

[1] Self-control is the ability to control one's passions and their influence over one's actions, and to stick to one's goals and plans in the face of temptation to do otherwise.

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_2

and plan their execution. These include minimal rationality, the ability to comprehend the means required to realize his goals, the mental faculties necessary to plan actions, etc." (Raz 1986, pp. 372–373). Raz also asserts that "[R]eason affects our choice of ends…just as much as it affects our deliberations" (Raz 1999, p. 73). Raz emphasizes what can be called the competence aspect of autonomy.[2] The autonomous person, as the author of his own life, must be competent at exercising the capacities for rational thought which Raz enumerates. Raz's detailed account, however, focuses exclusively on instrumental rationality: the ability of an agent to comprehend, deliberate about, and adopt means to his ends. While acknowledging that reason plays a role in the choice of ends, he does not address the question of how a rational agent should go about determining what ends to adopt.

According to Dworkin, the autonomous person's ends are self-chosen insofar as his intentions to pursue those goals are formed after critical reflection on his initial preferences, which he then either endorses or changes on the basis of that reflection. On this view, "[A]utonomy is conceived of as a second-order capacity of persons to reflect critically upon their first-order preferences…and the capacity to accept or attempt to change these in light of higher-order preferences and values" (Dworkin 1988, p. 20). Dworkin's view takes two aspects of autonomy into account. Like Raz, he recognizes the competence aspect of autonomy. Dworkin, however, does have something to say about the role of reason in choosing ends. An autonomous agent is one with the capacity for critical reflection on, and revision of, one's first-order preferences. This capacity is a prerequisite for the agent's autonomously forming intentions to pursue ends. Dworkin's view also recognizes a second aspect of autonomy, which we may call the authenticity aspect. According to Dworkin, an agent has achieved authenticity, in the sense that his preferences are truly his own, once he has brought his first-order preferences into alignment with his higher-order preferences. Authenticity is the result of successfully revising one's preferences; one achieves it by exercising the competence aspect of autonomy as well as one can. I will have something to say about authenticity at the end of Chap. 4.

What is the exact nature of the competence aspect of autonomy—the capacity to reflect critically on one's preferences, and then either endorse or revise them on the basis of that reflection? On this point, Dworkin's account is silent. The guiding thought of the next two chapters is that this capacity is a species of the capacity for practical reasoning. Specifically, it is the capacity for rational deliberation about what ends to adopt, deliberation which is conducted in the light of both the agent's persisting pre-deliberative attachments, and the evidence he gathers from his experiences of life and the world, which bear on the question of what ends would be best for him to adopt. My ultimate goal for the next two chapters is to develop a precise framework for representing competence in exercising this capacity of

[2] Raz also discusses another concept—"independence"—which he takes to be an aspect of autonomy. Independence, however, is not an aspect of autonomy, but a type of freedom—freedom from domination, whether physical or psychological, by others. Since I maintain a sharp distinction between autonomy and freedom, I do not discuss this concept in this part of the book. Independence is implied in my concept of freedom to exercise one's capabilities.

ends-deliberation. This framework will thus represent the core component of the capacity of autonomy. It is ideal competence in exercising this capacity that I will investigate; my framework will characterize the ideal of autonomy, rather than, to borrow Stanley Benn's term, the imperfect capacity of autarchy normally found in actual agents (Benn 1988). As in many other cases, modeling the ideal is the first step toward rigorous understanding, and once that step has been taken a model of the bounded capacity we ordinarily encounter may be possible.

Both Raz's and Dworkin's remarks on the value of autonomy also suggest that the core of the capacity of autonomy is a capacity for ends-deliberation (Dworkin 1988, p. 20; Raz 1986, p. 377). They both locate the value of autonomy in the fact that the autonomous agent gives meaning to his own life by choosing his own goals and commitments and constructing his life-plans for himself. But for these choices to count as the agent's way of constructing the meaning of his life, they must be reasoned choices. An agent cannot be said to determine for himself the meaning of his own life if these choices simply reflect preferences that the agent finds himself having. And it is hard to see how agents could reflectively endorse or revise their preferences in the way required for autonomy, if not as a result of a process of deliberating about what ends they should have for their lives and what options are worth choosing. I suspect that the difficulty of modeling reasoning about ends is responsible for the fact that so many authors with this general view of the nature and value of autonomy have failed to connect the capacity of autonomy with ends-deliberation.

Since the capacity of autonomy is, at least in its rational dimension, the capacity to adopt goals and construct plans to pursue them, it is natural to ask how a theory of autonomy such as mine relates to Michael Bratman's Planning Theory of practical reasoning. Much of the background for my framework for ends-deliberation is shared by the Planning Theory. Bratman describes the structure of plans as partial (insofar as the agent fills in the details of his plan as he proceeds) and as hierarchical (we plan to achieve some goals for the sake of achieving larger ones). This matches my discussion, in Chap. 4, of the nature of an ultimate end, the content of which is filled in as the agent selects the final ends which are constitutive of it. Bratman identifies two reasons for structuring plans in this way: the need to coordinate one's life subject to a limited capacity for reconsideration, and the need to cope with unforeseen events. I accept these as reasons for conceiving of one's ultimate end as an initially thin end that becomes more robust as its constituent ends are selected on the basis of deliberation. I also accept his constraints on plans as constraints on acceptable sets of final ends. These constraints are (1) internal consistency (it should be possible for one to achieve all of one's final ends); (2) strong consistency relative to one's (the beliefs that one's final ends are attainable and are at least as choiceworthy as the other available ends should be consistent with one's other beliefs); and (3) means-ends coherence (as time goes by, the agent should select final ends and acceptable means to those ends).

Bratman's central question is how to assess the rationality of forming an intention. Since intentions are the building blocks of plans, this question has a grander form: how to assess the rationality of the plans an agent makes for his life (Bratman

1987, p. 29). Bratman observes that one characteristic of intentions is that they set ends for further deliberation (Bratman 1987, p. 24). Deliberating about what intention to form, then, can involve deliberating about what end to adopt, not just deliberating about what means to an end to adopt. So Bratman's theory can be taken as an attempt to characterize the same competences in exercising rational capacities as my theory is concerned with—capacities that are integral components of autonomy. Bratman, and I, however, approach this question in very different—though complementary—ways.

The Planning Theory culminates in the Historical Principle of Deliberative Rationality. This principle states that:

> If A at t_1 forms the intention to ϕ at t_2 on the basis of deliberation at t_1, then it is rational of A at t_1 to intend to ϕ at t_2 iff:
>
> (1) for those intentions of A's that play a direct role as a background of A's deliberation, it is rational of A at t_1 so to intend; and
>
> (2) A reasonably supposes that ϕ is at least as well supported as its relevant, admissible alternatives. (Bratman 1987, p. 85)

My theory of the capacity of autonomy provides specific answers to the questions which this very general principle leaves unanswered. Each part of the principle presumes that it is possible to characterize the variety of practical reasoning that my theory is concerned with. The first part states that the intentions that serve as the background of a new deliberation must themselves be rational. The problem is that Bratman's theory is silent about how to assess the rationality of the agent's background intentions, and of the process of deliberation that leads from rational background intentions to a new rational intention. When the content of an intention is to adopt an end, a theory of ends-deliberation is required to assess the rationality of the intention. The second part of the principle states that the option settled on by the agent must be at least as well supported as the other relevant, admissible options. Again, whenever the agent adopts an end, the assessment of rational support will have to be made in the light of a theory like mine.

References

Benn, S. 1988. *A theory of freedom*. Cambridge: Cambridge University Press.

Bratman, M. 1987. *Intention, plans and practical reason*. Cambridge: Cambridge University Press.

Christman, J. 1991. Liberalism and individual positive freedom. *Ethics* 101(3): 343–359.

Dworkin, G. 1988. *The theory and practice of autonomy*. Cambridge: Cambridge University Press.

Frankfurt, H. 1988. Freedom of the will and the concept of a person. In *The importance of what we care about*, 11–25. Cambridge: Cambridge University Press.

Mele, A. 1995. *Autonomous agents: From self-control to autonomy*. New York: Oxford University Press.

Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.

Raz, J. 1999. Explaining normativity: On rationality and the justification of reason. In *Engaging reason*, 67–89. Oxford: Oxford University Press.

Thalberg, I. 1989. Hierarchical analyses of unfree action. In *The inner citadel essays on individual autonomy*, ed. J. Christman, 123–136. New York: Oxford University Press.

Chapter 3
Autonomy and Practical Reasoning

1 Introduction

My goal in these two chapters is to develop a philosophically satisfying formal dynamic account of rational deliberation about ends. I take the Aristotelian view that the core of the rational dimension of autonomy is the capacity for excellent deliberation about ends. So what follows will serve as my detailed characterization of (an idealized version of) the rational dimension of the capacity of autonomy. The account of autonomy will be completed in Chap. 7, when I address the capacity for self-control. Chapters 6 and 7 also contain a detailed, neo-Aristotelian characterization of individual freedom. Together, these two accounts—of autonomy and freedom—constitute my account of individual liberty. Most philosophical work on ends-deliberation has focused on the question of whether any such thing is possible, with arguments for and against its possibility being advanced against each other.[1] Rather than entering into this debate, I will simply present a consistent and coherent picture of what I believe ideally rational ends-deliberation looks like. This presentation will obviate any need for an argument in favor of the mere possibility of this type of deliberation.

We must be exceptionally clear at the outset about what is meant here by a formal dynamic account of ends-deliberation. In what follows, I will be adapting and applying the formal apparatus of modern decision theory to represent the steps taken by an ideally rational agent who has preferences over ends, but does not initially endorse those preferences reflectively, in order to arrive at the adoption of a set of preferences which he does reflectively endorse. My aim is thus a precise representation of the exercise of autonomy, as I understand it. This is a very different sort of task from the one undertaken in constructing a traditional decision-theoretic model. Most significantly, formulating an original set of axioms of rationality and proving a representation theorem from those axioms—the very heart of constructing a decision-theoretic model—is no part of the task I aim to

[1] On the history of this debate, see (Richardson 1986, ch. 1).

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_3

achieve here. Rather, my task consists in using an existing decision-theoretic model to precisely articulate the process of deliberating about ends from the perspective of an ideally rational agent—an agent whose preferences satisfy the constraints of the decision-theoretic model, and thus can be adequately represented in the formal language of expected value theory. My account of autonomy begins with this background in place, and uses this formal language to provide a precise and consistent description of each stage on the agent's path from unreflective to reflective preferences.

Given that such a description is the goal, we can see why the present task lies outside the scope of decision theory proper, and why the method of decision theory—axiomatization followed by proof of a representation theorem—is inapplicable here. A representation theorem can show us that given a certain set of constraints on an agent's meta-preference over possible preference-rankings, the agent's meta-preference can be represented by assigning a numerical value to each preference-ranking that appears in it. As we will see, we can use existing decision-theoretic models to represent just such a meta-preference, and this is where we must begin an account of the process of ends-deliberation. But the decision-theoretic model itself can offer us no guarantee that the agent reflectively endorses his meta-preference. No additional structural constraints (axioms) on the meta-preference will guarantee reflective endorsement. To reflectively endorse one preference rather than another is something more than to assign a higher value to it—this follows from the fact that we can value unreflectively. Reflective endorsement can only be secured as the end result of a process of adjusting one's preferences in the light of evidence about what one takes to be choiceworthy. To give an account of ends-deliberation is to illustrate this process. The procedural approach to rational ends-deliberation taken here has some important similarities to the procedural approach to rational instrumental deliberation developed by Brian Skyrms (1990). Indeed, Skyrms' work is the closest to being a precursor of the model I will present, and in the next chapter I will discuss the similarities between his model and my own, as well as the significant differences, which are primarily due to the shift in context from instrumental to ends-deliberation.

These two chapters are thus an exercise in applying decision theory to answer the philosophically important question of how exactly we should understand individual autonomy. They are not a piece of work in decision theory proper, but instead an example of how moral and political philosophy can use decision theory to give better answers to the questions that lie within their purview. Indeed, I believe that this is a question which only philosophy can answer, and careful philosophical consideration must guide our application of decision theory to this problem at every stage. The formal account developed here will only succeed if it presents us with a plausible and defensible picture of how we ought to understand the concept of individual autonomy, at least in an idealized context.

Although our focus is on ends-deliberation, we cannot fully understand autonomy without understanding excellence in instrumental deliberation as well. An autonomous agent must be skilled not only in choosing ends, but in choosing his

course of action for pursuing his ends. Accordingly, I begin the chapter with a brief discussion of the standard philosophical view of instrumental reasoning, as recently articulated by Robert Audi. I find it lacking in several respects. I then discuss the attempts of Henry Richardson and Elijah Millgram to develop informal, philosophical theories of reasoning about ends (Millgram 1997; Richardson 1986). Richardson and Millgram are the two philosophers who have self-consciously set themselves the task of working out a theory of ends-deliberation. Though neither succeeds, both contribute important insights which I will incorporate into my own theory. The next section of the chapter introduces the apparatus of modern decision theory, which provides a more satisfactory way to represent instrumental reasoning, and which I will make use of in the next chapter in developing my model of reasoning about ends. For reasons I discuss below, I rely primarily on the work of Frank Ramsey, particularly as it has been augmented by the research of David Sobel and Richard Bradley. It is Ramsey's version of decision theory that provides the best foundation for a dynamic model of ends-deliberation. The final section examines some recent work in the decision theory literature which is relevant to the primary task at hand. I discuss work by Amartya Sen, David Kreps, and Franz Dietrich and Christian List. The writings of each of these authors contain important contributions to the background against which I will develop my model. Exploring their work will allow me to bring into sharp focus exactly what it would mean to provide a precise and philosophically satisfying model of ends-deliberation.

2 The Standard Philosophical Model of Instrumental Reasoning

In a recent book, Robert Audi offers the following basic schema for practical reasoning:

> Major premise – the motivational premise: I want ϕ;
> Minor premise – the cognitive (instrumental) premise: My *A*-ing would contribute to realizing ϕ;
> Conclusion – the practical judgment: I should *A*. (Audi 2006, p. 96)

Audi notes that the term "want" in the major premise must be understood "in the broad sense of 'want' encompassing any kind of motivation, intrinsic or extrinsic" (Audi 2006, p. 92) and that in the minor premise "'contribute to' must be taken to encompass many kinds of beliefs, including…constitutive means-ends beliefs" (Audi 2006, p. 97). Audi's schema is a useful way to represent the most popular philosophical view of instrumental practical reasoning, and I will refer to it as part of the "standard model" of instrumental reasoning. The process it represents is reasoning insofar as it is the drawing of a conclusion from premises; it is practical insofar as its conclusion concerns what the agent should do; and it is instrumental insofar as the action recommended by the conclusion is a means, and is recommended as a means, to the agent's stated goal.

Let us see what we can learn from the standard model about the process of reasoning that leads the agent to believe that *A*-ing would contribute to his realizing ϕ, i.e. the process of reasoning about what means to select. On this point the model is not entirely silent. Audi offers a number of variants of the basic schema that at least provide us with some additional information about means. The first is a necessary condition schema, which Audi believes has the following valid form:

> I have an overriding want to ϕ;
> *A* is necessary to realizing ϕ, and is the only unrealized necessary condition for ϕ;
> So I should, all-things-considered, *A*. (Audi 2006, p. 141)

Also valid, according to Audi, is the optimality schema, in which the major premise is the same, and the minor premise expresses that *A* is the best way to realize ϕ (Audi 2006, p. 141). If we replace the minor premise with the assertion that *A* is sufficient for realizing ϕ, the only conclusion that follows is that I have some reason to *A*, since some other means might be preferable (Audi 2006, p. 141). Audi notes that in many (if not most) cases of practical reasoning, the agent is in a position in which he needs to determine what to do, not merely what he has some reason to do, and so he draws a conclusion to this effect which does not follow deductively from his premises. In such cases, we can evaluate how reasonable the inference is, where reasonableness is understood in terms of whether we are warranted in believing that the conclusion is more likely to be right than not (Audi 2006, p. 142).[2] He offers three grades for reasonableness. A piece of practical reasoning exhibits minimal adequacy if the conclusion is no less likely to be true than false, given the premises; standard adequacy if it is more likely to be true; and cogency if the degree to which it is more likely to be true exceeds that required by standard adequacy (Audi 2006, p. 143).

So the standard model has a little to tell us about what sorts of means figure in practical inferences (necessary, best, etc.) and about the nature of the inferences in which those different sorts of means appear (how to make them valid, when they are invalid but reasonable, etc.) But not much has been said about how to deliberate about selecting means. All we can yet assert is that if there is some available means whose selection we recognize will make our practical inference valid, we should choose that means; and if not, we should choose the one that will make the inference as reasonable as possible. Perhaps no deliberation is required to select the means that will render the necessary condition schema valid; if such a means is available, we either recognize that it is the final necessary ingredient for guaranteeing the achievement of our end or we do not. But we should want more than this. Studying instrumental reasoning should enable us to say something about how to make a

[2] Audi refuses to understand reasonableness in terms of how probable it is that the conclusion is correct; but his insistence that probabilities cannot be assigned here stems from a misunderstanding of the nature of such a probability assignment. He seems to have in mind our frequent inability to determine "objective" probabilities for the potential outcomes of a choice situation. But the existence of objective probabilities is unnecessary to a probabilistic description of what he calls reasonable inferences. Subjective probabilities, elicited from preferences via Ramsey's method of offering wagers, are all that is needed to model decisions made under uncertainty. I outline Ramsey's theory below.

reasoned selection between two (or more) means in a much wider range of cases. For that purpose, we will have to turn to decision theory. Before we do, let us examine two recent attempts to develop an informal, philosophical theory of reasoning about ends.

3 Specification and Practical Induction

Richardson and Millgram have made the only recent attempts at developing a theory of ends-deliberation. In developing my theory, I wish to take advantage of the work they have already done and adopt whatever in their theories will aid my own project.

I begin with Richardson. His central claim is that deliberation about ends proceeds by what he calls *specificational* reasoning, a type of reasoning distinct from means-ends reasoning. He begins by introducing the assertion that, given some valuable end q, most actions necessary to achieving q ought to be done. Specificational reasoning takes us from that assertion to the assertion that most actions necessary to achieving some end p should be done, where the latter assertion (regarding end p) is a specification of the former (regarding end q). Richardson calls these assertions *end-norms*. Every end-norm has what he calls an absolute counter-part, which is formed by replacing "most actions" in the original norm with "all actions" (Richardson 1986, p. 72). End-norm p then counts as a specification of end-norm q iff: (a) every instance of the abs.-counterpart of p counts as an instance of the abs.-counterpart of q; (b) end-norm p contains a description of end p which is one possible narrower or more specific description of end q, or asserts where, when, why, how, by whom or for whom end p (which may be the same as end q) is to be pursued, where these details are absent in end-norm q; (c) none of these additions to the original end-norm found in the specified end-norm is substantively irrelevant to the original end-norm (Richardson 1986, p. 73).

Richardson's theory is complex, and he elaborates it in great detail. But the essential points are as follows. We use specificational reasoning to deliberate about which ends to adopt and pursue in three different practical contexts. First, we use it to promote merely instrumental means to final ends. Richardson gives the example of a politician who identifies helping the homeless as an effective means to winning re-election. At first, he sees this as a merely instrumental means, and so he decides he will help the homeless provided that there are reporters around, that the situation is not too unpleasant, and that he gets a boost in the polls as a result. He also plans to stop once elected. When he begins to help the homeless, however, he is suddenly disgusted with himself for failing to see that activity as valuable in its own right. He then decides that he will help regardless of who is watching, no matter how unpleasant the situation may be, and that he will continue helping whether or not he is elected (Richardson 1986, pp. 84–85).

The politician in Richardson's example begins by viewing an action as a merely instrumental means, realizes he should not view it in that way, and starts viewing it

as a final end. He uses specificational reasoning to move from his first, highly restrictive specification of the norm "do most actions necessary to help the homeless," to a much less restrictive specification that reflects his new view of that action as a final end. But it seems that Richardson has simply passed over the most important aspect of this example. If what we want is a theory of ends-deliberation, then the question we need to answer is: how did the politician arrive at the conclusion that he was wrong to view helping the homeless as a merely instrumental means to election, and that it is in fact an end in itself? All Richardson has to tell us is that the politician suddenly felt appalled with himself, and then he knew he should value helping the homeless as an end in itself (emphasis added):

> [T]he thought of his own machinations *suddenly disgusts* him…These new feelings are a sign that he *has already learned* something from this little deliberative experiment…No, he decides, he will refuse to specify his involvement with the homeless so as to maximize his chances of getting elected [emphases added]. (Richardson 1986, p. 84)

But it is precisely that process of moving from one belief to the other, in which his observation of his feeling of self-disgust certainly plays a role, that we should be trying to represent precisely and explicitly in a theory of ends-deliberation. Richardson promises to explain in greater detail what he means when he says that the politician's feelings signify he has already learned something. But when he gets around to doing so, all he tells us is that "The medium of [the] concrete reflective self-awareness of ends, often, is emotion. Our emotions generally express our normative commitments" (Richardson 1986, p. 186). Whatever process resulted in the change in the politician's preferences remains obscure; we are simply told that he became fully aware of this change by observing an emotional reaction.

In this example, then, specification is not the process of reasoning whereby an agent determines that his beliefs regarding what is choiceworthy are wrong, and then arrives at a new judgment about what is choiceworthy. Rather, specification kicks in after the agent has already re-ordered his preference-ranking. Given that the politician now prefers getting elected and helping the homeless to getting elected and not helping them, he needs to figure out the best means to helping them from among a wide set of available means to helping them, rather than from among the much smaller set of means to helping them that will also increase his chances of winning the election. In this example, then, specificational reasoning is the process by which the politician determines how to expand the set of means to helping the homeless from which he will choose. This is quite different from deliberating about ends.

Richardson's second role for specification in ends-deliberation is the dissolution of conflict between ends. He argues that when ends conflict, we need not choose between them without altering them, and thus be forced to decide that one is more important than the others. We can specify one or more of the conflicting ends in such a way that the conflict dissolves (Richardson 1986, p. 171). The problem, which Richardson acknowledges, is that there may be several ways of doing this, and we need some way of choosing one specification (or set of specifications) over the others. His answer is to choose the specification that best coheres with one's

other ends which are not currently under consideration (Richardson 1986, p. 173). I will have something to say about the notion of practical coherence in the course of developing my own account. For now, we should note that once again, specification is not playing the role of the sort of reasoning we engage in when we deliberate about ends. This time, it is a precursor to that deliberation. It provides the agent with a set of alternately specified groups of ends, from which he must then choose. The process of engaging in specificational reasoning in this content, therefore, is itself a determinate means in Aristotle's sense—it is an act of determining more precisely what the end which the agent might go on to pursue consists in. But it is how the agent chooses between these ends that is the concern of a theory of ends-deliberation.

The third role for specification involves sets of settled, non-conflicting ends. Richardson claims that we use specificational reasoning in determining which ends are pursued for the sake of which other ends—that is, in order to work out the hierarchical structure of our set of ends (Richardson 1986, pp. 174–178). If an agent has two final ends which are each a means to the other, then deciding that one is pursued for the sake of the other is a way of further specifying both of those ends. But again, the crucial question for a theory of ends-deliberation is how the agent goes about determining which end is pursued for the sake of which. Richardson's answer in this case also invokes coherence—it may cohere better with one's other ends and the way they are structured to say that *this end* is pursued for the sake of *that one*, rather than vice-versa (Richardson 1986, p. 179). As I said, I will return to the possible role of the notion of practical coherence in a theory of ends-deliberation. For now, we should note that in this case there does not seem to be any specificational reasoning whatsoever. The agent somehow deliberates about which of these two ends best coheres with his other ends (without further specifying the content of any of his ends), and simply in virtue of reaching a conclusion in that deliberation, finds that one of those two ends has been further specified just insofar as it is now understood as being pursued for the sake of the other (Richardson counts "X is to be pursued for the sake of Y" as a specification of "X is to be pursued.") The question of how he deliberates about these ends has once again been passed over.

Richardson's specificational reasoning is not the sort of reasoning an agent employs when he deliberates about ends. Rather, it is the sort of reasoning he employs after he engages in ends-deliberation resulting in the promotion of a new final end, in order to determine how to expand the choice set of available means to achieving the newly promoted end; and it is the sort of reasoning he employs before he engages in ends-deliberation to resolve conflicts between ends. Deliberating about which ends are pursued for the sake of which others results in ends that are further specified, but that specificational reasoning does not seem to play a role in this process. Rather, the further specified ends are the by-products of reasoning about the practical coherence of ends. Let us proceed, then, to Millgram's theory, which he calls Practical Induction. As we will see, this theory offers a couple of basic insights that are crucial for developing a more exact theory of ends-deliberation.

Millgram defines practical induction as the type of reasoning we engage in when we determine what is choiceworthy based on our particular lived experiences (Millgram 1997, p. 81). Such reasoning requires premises, and the premises will have to take the form of practically relevant observations. But are there any such observations? He explains that "the sense in which we need to show that practical observations can be taken seriously is just this: there are practical judgments of particulars, formed in response to, and reflecting, experience, that can be legitimately used as premises in practical inductions" (Millgram 1997, p. 106). Millgram then focuses his discussion of practical observation on observations of pleasure, which he understands as the feeling that things are going well, and which he takes to be an indication that things really are going well (Millgram 1997, p. 117).

From this short summary of Millgram's view, we can extract two key ideas. The first is that we arrive at conclusions regarding what is choiceworthy on the basis of experience, and the second is that the relevant experience takes the form of observations of *evidence* that support claims of the form that one available option is more choiceworthy than another. By introducing the idea that we can determine what is choiceworthy based on experience, he widens the horizon of a theory of ends-deliberation well beyond where it was set by Richardson's appeals to practical coherence.

Milgram's theory has a number of important flaws. He fails to appreciate the variety of considerations that may play the role of evidence for an end's choiceworthiness—a point explored below—and instead focuses solely on observations of feelings of pleasure. He spends most of his book arguing that this sort of reasoning is something that must be possible, and that we must engage in, rather than developing a precise account of it. And his arguments, like Richardson's, mostly take the form of vignettes meant to illustrate someone engaging in this type of reasoning. These vignettes are never analyzed with sufficient care to reveal the structure of the deliberations they are meant to illustrate. He describes, for example, a case of an agent deciding to adopt a new final end, and a case of an agent learning from experience that he should change the position of one of her ends in her hierarchy of ends (Millgram 1997, pp. 108 & 111). But in his analysis of these cases, he fails to distinguish them from each other, or from other cases he describes in which an agent learns which means to an end is best, or in which an agent resolves a conflict of ends by choosing one specified end rather than another (Millgram 1997, pp. 62 & 59). These distinct types of cases are all lumped together by Millgram under the heading of learning what is choiceworthy from experience. And he never describes the steps in reasoning that the agent proceeds through in between his experiences and his new normative conclusions. Rather, he repeatedly describes his agents as simply realizing these conclusions straightaway, once they have articulated their observations of their own experiences. Here are a couple of typical examples (emphasis added):

1. Diana, who works maintaining indoor plants in offices, is trying to revive a dying tree. Recently, she has been receiving compliments on the condition of the other plants in the office. "[T]he care she was putting into the tree was carrying over into the rest of the account, but she herself did not *realize* that this was happening

until the compliments began…She *has learned* that she likes challenges that involve improving and reviving living things." (Millgram 1997, p. 108)

2. Ellen has been working nights as a waitress to support the pursuit of a career as a dancer. "[S]he *found herself* ever more unable to cope with the day-to-day details of living…Although she felt miserable, it took her a long time until she *understood* what was making her feel that way…At that point, she *was resolved* not to waitress any more. She now…dances [just for fun] in the evenings." (Millgram 1997, p. 110–111)

The mechanics, as it were, of moving from the observation that one feels better when doing X than when doing Y, to the conclusion that X is more worthy of choice than Y, are left mysterious. To be sure, each agent arrives at the conclusion we would expect given the experience Millgram describes him as having; but nothing detailed is said about the structure of the agent's process of reasoning his way to the conclusion.

Millgram's work, then, leaves all the heavy-lifting of a theory of ends-deliberation still to be done. We will see, however, that despite these shortcomings, something like the notions of practical induction and observation that make up his view's core will prove to be key components in a satisfactory account of ends-deliberation.

4 Decision Theory: Ramsey and Beyond

We have found good reason to be dissatisfied with the standard informal philosophical account of instrumental reasoning, and recent attempts to provide an informal philosophical account of reasoning about ends. We now turn to decision theory as a formal theory of instrumental reasoning. This will provide us with most of the background we need to develop a formal theory of reasoning about ends.

4.1 Preferences, Preference-Rankings, and Valuations

Bayesian decision theory is designed to represent the instrumental reasoning of a rational agent whose ends are given. In the next chapter, I will be using the framework of modern Bayesian decision theory to develop my model of rational ends-deliberation. So I must speak of my deliberating agent in terms of his preferences, his subjective probability judgments, and his expected value judgments. Some background in the content of these notions and the implications of appealing to them is needed in order to steer clear of a number of philosophical confusions. The first important point is that a preference is nothing like a desire in the philosophical sense of that term. Preferences are judgments about what is better than what. The economist and philosopher Daniel Hausman has recently argued that preferences should be understood as *total comparative evaluations* (Hausman 2011, p. 3). That

is, he understands a preference for *a* over *b* as a judgment that *a* is superior to *b* with respect to every consideration taken by the agent to be relevant to the judgment. This is close to being right. By the end of the next chapter, I will have developed an account of how a rational agent deliberates over and intentionally adopts and revises his preferences. I reserve the understanding of preferences as total comparative evaluations for preferences that have been adopted as the result of excellent deliberation. We can understand other preferences as aiming to be evaluations of this sort, but falling short. The agent who does not deliberate excellently about his preferences may fail to end up with preferences that reflect every consideration he takes to be relevant in the way he takes it to be relevant.[3] We can think of preferences as ranging over actions, their outcomes, and the states of affairs in which we act. However, in the formal decision theory I will employ, preferences are strictly understood within the theory as ranging over propositions which describe the performance of actions, the realization of outcomes, and the obtaining of states of affairs. Thus, in the assertion that *a* is preferred to *b*, "*a*" and "*b*" name propositions, and what is preferred is the truth of proposition *a* to the truth of proposition *b*.[4] Under normal circumstances, an agent's preferences determine his intentions—the plans he makes about what he is going to do—and his choices—the actual actions he performs (Hausman 2011, p. 2). We can interpret "better than" in any number of more specific ways—in the sense of being of greater value, or more fitting or appropriate, or more worth choosing, or more welcome.[5] Which interpretation is best may vary with context, though I will most often speak in terms of the choiceworthiness of performing an action, and the welcomeness of an outcome. So to prefer *a* to *b* is to believe that *a* is better in one of these senses than *b*.[6]

[3] This point will turn out to be important to explaining the possibility of failures of self-control (i.e. weakness of will).

[4] Richard Jeffrey first argued that we should take preferences to be over propositions (Jeffrey 1990). For additional argument in favor of this approach, see (Joyce 1998).

[5] Recently, a few philosophers have resurrected and defended the view that a desire for something just is a belief that that thing is desirable. See in particular (Bradley and List 2009), which responds to an important criticism of this thesis made by David Lewis. The most famous argument for distinguishing beliefs from desires is that these two attitudes have different "directions of fit": beliefs aim to fit the world, whereas we aim to make the world fit with our desires. But distinguishing the two attitudes on this basis proves to be harder than it first appears (Sobel and Copp 2001). Even if we accept the desire-as-belief thesis, we can only identify desires and preferences if we identify the belief that X is desirable with the belief that X is choiceworthy. This may be perfectly acceptable, but it is a point I wish to remain agnostic about.

[6] I will simply assume that preferences are beliefs of this kind. This is arguably the view of preference taken by Kenneth Arrow, one of the founding fathers of social choice theory (Arrow 1951, p. 23). For some interesting remarks supporting this position, see (Broome 2006, pp. 206–208). And for a sustained argument from a Sellarsian perspective against desire/preference non-cognitivism, with which I am largely in agreement, see (J Heath 2008, ch. 4–5). For an empirically based argument against the very existence of desires as they are usually understood in modern philosophy (outside the desire-as-belief camp), to which I am sympathetic, see (DeLancey 2002, pp. 3–12). Modern economic theory is based on *revealed* preferences—that is, an agent is taken to prefer *a* to *b* iff, given a choice between *a* and *b*, he chooses *a*. The agent's preferences are thus inferred from his choices. The orthodox position is that this sort of behaviorist interpretations is (at least) appro-

Preferences are subjective insofar as different agents may have different judgments of this sort, and be led to these different judgments by their different circumstances. Given the particular situation of one agent as opposed to another, moreover, it may be true that one option is more choiceworthy for one agent but less choiceworthy for another. More appropriate than the language of subjectivity and objectivity is Sen's terminology of *position-dependence* and *author-invariance* (Sen 1985, pp. 183–184). What is most choiceworthy for an agent depends on the position, the particular circumstances, of that agent. These include his particular tastes and interests. But there is still a fact of the matter about what is worth choosing for an agent, given all the particularities of his position; this is the author-invariance of choiceworthiness.

A valuation function is a real-valued function that represents an agent's preferences numerically.[7] Any strict partial ordering can be represented numerically. Given any set of alternatives, if it is true of any two members of that set a and b that either $a \succeq b$ or $b \succeq a$, and that $a \succeq b$ and $b \succeq c$ implies $a \succeq c$, then we can assert that $a \succeq b \leftrightarrow f(a) \geq f(b)$ for some function f (Fischburn 1972, p. 3). For an agent's preferences to be representable in terms of a cardinal valuation function those preferences need only satisfy a few additional requirements. The simplest such set of requirements is the one formulated by John von Neumann and Oskar Morgenstern (von Neumann and Morgenstern 1944). The von Neumann-Morgenstern requirements are stated in terms of preferences over lotteries, where a lottery L is a set of options over which the agent has preferences, with a probability associated with attaining each option within the lottery. The conditions are:

1. Completeness: For all lotteries L and M either $L \succeq M$ or $M \succeq L$
2. Transitivity: If $L \succeq M$ and $M \succeq N$ then $L \succeq N$
3. Continuity: If $L \succeq M \succeq N$ then there is a $0 \leq p \leq 1$ such that $p(L) + (1-p)(N) \sim M$
4. Independence: If $L \sim M$ then $p(L) + (1-p)(N) \sim p(M) + (1-p)(N)$

If these requirements are met, the agent's preferences can be numerically represented by the values of a valuation function such that $a \succeq b \leftrightarrow v(a) \geq v(b)$. This function is cardinal insofar as differences in magnitude between values of the function are significant. Thus we can make assertions of the form $v(a) - v(b) > v(c) - v(d)$.

We should note two philosophically important points about the process of constructing a valuation function. The first is that although these values are often given a hedonist, or utilitarian, or at least consequentialist interpretation, they need not be. In a given decision problem, outcomes can be ranked according to the extent to which they will be achieved virtuously, or through responsiveness to reasons, or by

priate for the science of economics (since it is the choice behavior of consumers that the economist is able to observe and record), but it has been forcefully criticized by Amartya Sen (Sen 1999).

[7] These functions are more often called "utility functions," and their values called "utility numbers" or "utility values." But I avoid use of the word "utility," since it inescapably evokes associations with utilitarianism. I thus follow Sen in speaking of valuation functions and their valuations.

conforming to the moral law, etc.[8] Available actions can accordingly be ranked according to how well supported by reasons they are, or how virtuous they are, or how well they conform to the moral law, etc. Whether we conceive of the agent as ranking actions in virtue of the values of their outcomes, or ranking outcomes in virtue of the goodness (or rightness, or virtuousness, or what have you) of the actions that produce them, is an open interpretive question, and a free interpretive choice. This issue is orthogonal to the formal apparatus of], which we use to represent preferences in a mathematically convenient way. We can take the agent as basing his preferences on whatever normative foundation he likes, or on none at all—his judgments about what is more choiceworthy than what may be based on nothing more than what happens to strike him as more choiceworthy.

The second point is that all we have to assume in order to construct a cardinal valuation function is that the items in the preference-ranking are *comparable*, and that those comparisons satisfy the requirements of a representation theorem (such as the von Neumann-Morgenstern axioms). We did not have to assume that the agent's preferences are "commensurable," either in the sense that there exists some unique, universal currency which expresses precisely how much of some single fundamental value each item in the ranking contains, or in the sense that all of the agent's preferences are jointly realizable.[9] Again, in constructing a valuation function, we did not need to assume that acts or outcomes were ranked according to how pleasant they are, or how happiness-conducive, or what have you. So there is no reason to take these values to be numbers of "happiness points." A cardinal valuation function simply represents numerically a set of ordinal preferences that satisfy the requirements of a representation theorem. Nor does representing preferences in terms of numerical valuations commit us to a "compositional" view of goodness.[10] Since the valuation function is not tracking the amount of any fundamental value present in all the items in the agent's preference-ranking, there is no reason to assume that more—in the quantitative sense—will always be better. Though the agent may prefer eating a scoop of ice-cream to eating a slice of cake, it need not be the case that

[8] In L.J. Savage's formulation of expected utility theory, the third postulate requires that preferences over consequences be act-independent. If one outcome is preferred to another, it must be preferred regardless of the act through which either outcome would be obtained. This does not, however, undermine non-consequentialist interpretations even of his construction of expected utility functions. The set of outcomes is an arbitrary subset of the set of possible states of the world. So nothing stops us, for example, from using a fine-grained division of outcomes, so that, for example, obtaining good x honestly is a different outcome from obtaining good x dishonestly. Act-independence would then only require that if obtaining good x honestly is preferred to obtaining good y honestly, it is so preferred regardless of what *honest* acts would lead to either outcome. And our division of outcomes can be even more fine-grained if we like. A choice set with this sort of fine-grained division contains what Sen refers to as "comprehensive outcomes" as opposed to the coarser "culmination outcomes" (Sen 2002, p. 595).

[9] Isaiah Berlin uses "commensurability" in this sense. See (Berlin 1998).

[10] Michael Stocker argues at length against maximizing theories of decision which do endorse a compositional view of goodness. (Stocker 1990, pp. 281–309). He then admits, however, that modern utility theory is not committed to this view (Stocker 1990, p. 310).

he prefers eating 19 scoops of ice-cream to eating a slice of cake.[11] Though it is true that a higher valuation is always better, this means no more than that what is more choiceworthy—higher in the preference-ranking—is always better than what is less choiceworthy.[12]

The philosophical notion of commensurability, which has been cast as a sort of boogeyman in many discussions of reasoning and decision, is in fact so hopelessly imprecise and unnecessary that it should be abandoned (Anderson 1997; Lukes 1997; Richardson 1986). What most philosophers who worry about commensurability seem to associate with that term is what decision theorists study under the guises of the measurability and interpersonal comparability of a valuation function.[13] Measurability comes in levels—ordinal, cardinal, and so on, up to perfect measurability. A valuation function's level of measurability denotes the kind of information captured by the function's numerical representations of the agent's preferences. Ordinal valuation functions do no more than assign a greater number to a more preferred option and a lesser number to a less preferred one $\left(v\left(a\right)>v\left(b\right)\right)$. All ordinal valuations that do this are equivalent. Cardinal valuation functions provide information about preference intensity—if the difference between the valuations of a and b is greater than that between b and c on one cardinal scale, it will be so on all equivalent cardinal scales $\left(\left|v\left(a\right)-v\left(b\right)\right|>\left|v\left(b\right)-v\left(c\right)\right|\right)$. If value is perfectly measurable, then there is a unique function that assigns a unique number to each of a person's preferences, which uniquely and exactly represents the value of that alternative for that person. Interpersonal comparability also comes in degrees. Valuations may be non-comparable, in which case we have no information about whether one outcome is more valuable to one person than another. They may be unit-comparable—in which case we know whether the difference in the benefit to agent 1 of outcome a versus outcome b is greater than the same difference in benefit to agent 2 $\left(\left|v1(a)-v1(b)\right|>\left|v2(a)-v2(b)\right|\right)$; or level-comparable—in which case we know whether outcome a will leave agent 1 better off than it would leave agent 2 $(v_1(a)>v_2(a))$. Finally, valuations are fully comparable if they are both unit- and level-comparable. If a valuation function is perfectly measurable and fully comparable, there is a unique scale with which to represent every one of every agents'

[11] The assumption that utility always increases when more of a commodity is consumed is common in economic theory, but not essential to decision theory.

[12] Stocker's other argument against maximizing theories of decision is a defense of the claims that it is often not immoral to fail to do what is best, and that it is often implausible to assert that an agent who has decided against doing what is best is irrational (Stocker 1990, pp. 310–342). I certainly have no objection to the first claim. Decision theory is simply not concerned with, and not equipped to answer, the question of how much good one must do, in comparison with how much good one could do, in order to do what is morally required. I reject the second claim, but with the qualification that I do think rationality is a matter of degree. The agent Stocker is imagining has been irrational to *some* degree, but it may be a very slight one. And if it is slight enough, it would be implausible to think that anyone would be justified in actually *calling* him "irrational." But it is not implausible to maintain that he is, to some very slight degree, irrational.

[13] For an excellent discussion of the different levels of measurability and comparability of utility values, see (Barberà et al. 2004, pp. 1115–1123).

preferences. This sort of function thus captures the most information possible about agents' preferences. Arguing that valuation is perfectly measurable and fully comparable amounts to arguing that there is some single fundamental value, measured in some unique universal currency, in terms of which all alternatives must be evaluated. This is the likely target of philosophical worries about commensurability—they are worries about the strong value monism that it implies. But I am claiming nothing of the sort. In developing a framework for ends-deliberation, I will only assume that valuations of preferences are cardinally measurable. Cardinal valuation is not necessarily comparable across agents (and I do not assume that it is), and a non-comparable cardinal valuation function is unique only up to a positive affine transformation. There are an infinite number of valuation functions that represent a given agent's preferences equally well, assigning *different values* to those preferences. Using a cardinal valuation function to represent preferences simply *cannot* commit one to claims like "option *a* is exactly 2.34 happiness units better than option *b*"; that measure of distance between the values of the two options simply is not meaningful.

There is an objection looming. Perhaps the worry regarding commensurability is actually that the sort of comparisons among options the agent must make in order to have a preference-ranking that satisfies the requirements of a representation theorem would not be possible unless the agent were able to commensurate these options. It would then not be enough to point out that binary comparisons, suitably restricted, suffice for a cardinal representation. What would be needed is an argument that the existence of these comparisons themselves does not tacitly rely on the commensurability of the options being compared, in the sense that the value of those options can be precisely specified in some common coin. At the heart of this objection is the worry that one's preferences could not be anything but arbitrary unless one were able to commensurate one's options. If we were to reject the idea that we are able to commensurate our alternatives, then we would lose the ability to make non-arbitrary comparisons as well.

As Ruth Chang points out, however, the assumption behind this worry is that comparison must itself be a calculative process (Chang 1997, p. 18). It will only be the case that comparing alternatives presupposes that they are commensurable if the act of comparing them is the act of determining how much of some value, in a precise quantitative sense, they have. Why should we think this? Why should we assume that in judging one thing better than another, we must be able to say exactly how much better it is? One reason would be a commitment to a compositional view of goodness. If we think that if something is valuable, then more of it—in a quantitative sense—must be better, then we might be led to the conclusion that comparison is an essentially quantitative exercise. But we have good reason to reject a compositional view. Though I recognize the value of friendliness and prefer that a stranger who interviews me be fairly friendly rather than unfriendly, I also prefer that he be fairly friendly rather than *extremely* friendly (Chang 1997, p. 17). And once we reject the view that what makes one alternative more choiceworthy than another is its greater *quantity* of some value, it is hard to see why we should think of comparing as essentially quantitative. This holds even if we are committed to

representing preferences in terms of cardinal valuations. Such a representation does not imply that agents arrive at their preference judgments in the first place by assigning quantitative valuations to their available options. Rather, the point of the representation is that we can interpret the agent's decisions between items in his preference-ranking as if he were attempting to maximize the value of a function. As we are about to see, this sort of representation significantly enhances our ability to study the process of rational decision-making.

By committing myself to using cardinal valuation functions to represent agents' preferences, I am committing myself to the claim that the values of an agent's available alternatives are comparable—that it is possible for an (acceptably idealized) agent to have a complete, non-arbitrary preference-ranking.[14] The deepest objections to the use of valuation functions are thus objections to comparability itself. Chang does an excellent job of disarming the most popular objections to comparability one by one, until only an objection posed by Joseph Raz remains (Chang 1997, pp. 23–27). Raz's objection is as follows. There are some cases in which we cannot say either that one alternative is better than the other, or that they are equally good, *or* that if one were to be slightly improved it would then be better than the other. Suppose, for example, we were to try to compare the talent of Michelangelo with the talent of Mozart. Raz believes we would be right to resist the claim that either is better than the other, or that they are exactly as good as one another. Now suppose that there were an artist who was very nearly identical to Michelangelo, only slightly better than he. Call this artist Michelangelo^{+}. Although Michelangelo^{+} is more talented than Michelangelo, Raz claims that we would still be right to resist the claim that he is better, worse, or exactly as good as Mozart. So the case of Michelangelo and Mozart is, according to Raz, a case of genuine value incomparability.[15]

Chang does offer a response to Raz, which involves formulating a fourth comparative relation (alongside better than, worse than, and indifferent to): the relation of being *on a par with* (Chang 1997, pp. 25–27). Intriguing as this response is, I will not discuss it here. Rather, I maintain that Raz's claim, even if it is correct, does not impact my use of preference-rankings and valuation functions. For I need not claim that *all* values are comparable. In Raz's example, the type of value in question is aesthetic value—how good Mozart and Michelangelo are *qua* artists, or how good their works are *qua* works of art—and the comparative judgment the agent is attempting to make is an aesthetic one. I do not claim to offer a theory of aesthetic judgment, and I am happy to concede that in some cases, aesthetic values may be incomparable. My concern is with preference-rankings of potential actions and

[14] The comparability thesis is an old one. Aristotle seems to have held it: "…the kind of imagination concerned with deliberation is had only in those which are capable of reasoning, for deliberation whether to do this or that is already a task for reasoning, *and there must be one measure* (καὶ ἀνάγκη ἑνὶ μετρεῖν), for *one pursues what is superior*, hence one has the ability *to make one image out of many* [emphasis added]" (*De Anima* III.11 434a7-10).

[15] Raz refers to this as a case of incommensurability, but it should be clear that it is meant to be a case of incomparability as I am using the term.

ends; and an objection of the sort Raz makes does not seem to count against the claim that the values of an agent's potential actions are comparable. The relevant parallel case to the one Raz describes would be that of an agent who must choose between attending an exhibit of Michelangelo's work or a performance of Mozart's. Even if the aesthetic greatness of these two artists is incomparable, this fact does not render a reasoned preference between these two alternatives impossible. The possible bases for such a choice might include the frequency of opportunities to attend exhibits and performances in this agent's community; the agent's own interest in music and the visual arts; or the other factors contributing to the overall quality of the experiences such as the venues at which the events will take place. There is thus no intuitive force to the claim that an agent, taking account of the various particulars of his circumstances and interests, would be incapable of making a reasoned comparison and choice between two aesthetic experiences of incomparable aesthetic value.

We can of course recast Raz's objection so that it does apply to my assumptions. The objection would then assert the possibility of an agent being faced with a choice between two ends or courses of action, being unable to rank them with respect to each other, and still being unable to rank one of them against a slightly better (or worse) version of the other. Why should we insist on such a possibility? My hunch is that the motivating concern must be an epistemic one. There may be cases in which one cannot find any grounds for preferring one or the other of two options, but in which it just seems implausible to claim that each is exactly as good (or as bad) as the other. This is why the second part of the objection, involving the slightly better (or slightly worse) version of one option is important. If this second pair could be ranked, one would be pushed toward a judgment regarding the appropriate ranking of the first pair, however implausible that judgment may have initially seemed.

I believe, however, that I can accommodate the concern that fuels this objection while retaining my assumption of comparability. One key component of my framework for deliberation will be the ability of agents to assign probabilities to their preference judgments representing their credence in the correctness of those judgments. I can thus represent the sort of situation Raz has in mind as one in which the agent is completely undecided between the three possible preference judgments he could make: $p_1(a \succ b)$, $p_2(a \sim b)$, $p_3(b \succ a)$, where $p_1 = p_2 = p_3$. When the agent then considers an option a^+ slightly better than a, and compares it to b, his credence shifts in favor of the superiority of a^+, but not decisively; he now judges $p(a^+ \succ b) > p(b \succ a^+)$, but still judges $p(b \succ a^+) > 0$. Thus the agent remains uncertain whether a or a^+ is preferable to b. None of this points to incomparability; the agent is merely as yet uncertain and undecided about which comparison to endorse. If the epistemic concern which, I have conjectured, underlies Raz's objection can be satisfied in this way, we will need some other distinct and serious worry to tempt us away from an assumption as useful as comparability. Since I do not know of one, I am content to retain this assumption.

4.2 Valuation, Subjective Probability and Expected Value

Along with the controversial notion of cardinal valuation, Bayesian decision theory employs the perhaps equally controversial notion of subjective probability. The subjective interpretation of probability takes an agent's assignment of a probability to the truth of a proposition as a measure of that agent's degree of partial belief, or *credence*, that the proposition is indeed true. One of the great accomplishments of Frank Ramsey was to show that given certain requirements on an agent's preferences over outcomes (or rather over what he called *values*: sets of indifferent outcomes), not only can those preferences be represented by a valuation function, his qualitative judgments of the likelihood of those outcomes—judgments to the effect that one outcome seems more likely than another—can be represented quantitatively *via* a probability function (Ramsey 1931).

Ramsey begins by defining what he calls an "ethically neutral" proposition—one whose truth is believed with probability 1/2. X is an ethically neutral proposition if, and only if, an agent prefers outcome α to outcome β (or prefers any outcome indifferent to α to any outcome indifferent to β-which is to say any member of the value, or set of indifferent outcomes, $\underline{\alpha}$, to any member of the value $\underline{\beta}$), but is indifferent between the conditional gambles (1) α if X is true, β if X is false; and (2) α if X is false, β if X is true (Ramsey 1931, p. 178). The outcomes α and β may also be taken to be propositions. They describe possible future states of the world. More precisely, each outcome is a near-maximally complete proposition. That is, it is a proposition which only falls short of uniquely identifying a possible state of the world insofar as it remains consistent with both the truth and the falsehood of X.[16] In the first of these conditional gambles, the agent receives outcome α if X is true, and β if it is not. Ramsey then uses this definition to find a way to determine that the difference in valuation (or choiceworthiness) between α and β according to the agent is the same as the difference between two other outcomes γ and δ. Let X and Y be ethically neutral propositions. Suppose the agent is indifferent between the conditional gambles (1) α if X, δ if $-X$; and (2) β if X, γ if $-X$. Then the agent is also indifferent between the conditional gambles (3) α if Y, δ *if* $-Y$; and (4) β if Y, γ if $-Y$. The fact that the agent is indifferent between (1) and (2) and between (3) and (4) implies that the difference in valuation between α and β is the same as that between γ and δ. This is represented by $\alpha\beta = \gamma\delta$.

With these definitions of ethically neutral proposition, conditional gambles, and differences in valuation in place, Ramsey goes on to sketch a representation theorem. He famously, however, neglects to work out the details of his result. Recently, Richard Bradley has published a complete version of the proof of Ramsey's theorem (Bradley 2004). We assume that the agent's preferences over outcomes are restricted by the following 11 axioms:

R1. If P and Q are ethically neutral, and (α if P) (β if $-P$) $\succeq$ (γ if P) (δ if $-P$), then: (α if Q) (β if $-Q$) $\succeq$ (γ if Q) (δ if $-Q$)

[16] This is not how Ramsey himself understood outcomes. He thought of them simply as possible worlds. For outcomes as near-maximal world-propositions, see (Bradley 2004, p. 487).

R2. If (α if *P*) (δ if –*P*) ~ (β if *P*) (γ if –*P*), then:

(i) $\alpha \succ \beta$ iff $\gamma \succ \delta$
(ii) $\alpha \sim \beta$ iff $\gamma \sim \delta$.

R3. (i) $\phi \succeq \psi$ or $\psi \succ \phi$
(ii) If $\phi \succeq \psi$ and $\psi \succeq \omega$, then $\phi \succeq \omega$.

R4. If (α if *P*) (δ if –*P*) ≽ (β if *P*) (γ if –*P*) and (γ if *P*) (ζ if –*P*) ≽ (δ if *P*) (η if –*P*), then:
(α if *P*) (ζ if –*P*) ≽ (β if *P*) (η if –*P*).

R5. $\forall(\alpha,\beta,\gamma)\left[\exists(\delta):(\alpha ifP)(\gamma if-P)\sim(\delta ifP)(\beta if-P)\right]$.

R6. $\forall(\alpha,\beta)\left[\exists(\delta):(\alpha ifP)(\beta if-P)\sim(\delta ifP)(\delta if-P)\right]$.

R7. For every conditional prospect (α if *X*) (β if –*X*) there exists a world γ such that:
(α if *X*) (β if –*X*) ~ γ

R8. Any non-empty set of values which has an upper bound has a lowest upper bound.[17]

R9. α ≽ β iff α ≽ (α if *P*) (β if –*P*) ≽ β.

R10. Let{P_1, P_2,…,P_n} be a partition of propositions. Then:
$\forall(\gamma, \delta, \ldots, \beta), \exists(\alpha: (\gamma \text{ if } P_1)(\delta \text{ if } P_2)\ldots(\beta \text{ if } P_n) \sim (\alpha \text{ if } P_1 \cup P_2)\ldots(\beta \text{ if } P_n)$

R11. For every *X* which is not a logical truth or falsehood, there is a possible world with any assigned value in which *X* is true, and one in which *X* is false.

With these axioms assumed, Bradley proceeds to prove the representation theorem. The details of the proof need not concern us here. In brief, Bradley invokes the measurement theorem of the nineteenth century German mathematician Hölder, which defines a set of mathematical objects called "algebraic difference structures," and then shows that given the Ramsey-axioms on preference, the pair <$\underline{\Gamma}$x $\underline{\Gamma}$, ≽>, where $\underline{\Gamma}$ is the set of all values, is an algebraic difference structure. It follows from this result that:

Existence: There exists a valuation function *v* on the set of values $\underline{\Gamma}$ such that for all $\underline{\alpha},\underline{\beta},\underline{\gamma},\underline{\delta} \in \underline{\gamma}$: $\underline{\alpha} - \underline{\beta} \geq \underline{\gamma} - \underline{\delta}$ iff $v(\underline{\alpha}) - v(\underline{\beta}) \geq v(\underline{\gamma}) - v(\underline{\delta})$.

Uniqueness: If *v'* is another such valuation function, then there exist $a,b \in R$ such that $a>0$ and $v' = av + b$.

And it follows from Existence that: for all $\underline{\alpha},\underline{\beta} \in \underline{\gamma}$: $\underline{\alpha} \geq \underline{\beta}$ iff $v(\underline{\alpha}) \geq v(\underline{\beta})$. $v(\cdot)$ is therefore a valuation function of the familiar sort, and is unique up to a positive affine transformation.

We can then use valuations of outcomes to define the agent's degree of partial belief, or credence, in a proposition *X*. Suppose that outcome α is consistent with the truth of *X* and outcome β is consistent with the truth of –*X*, and that there is an outcome ζ such that (i) it is not the case that α ~ β; and (ii) ζ ~ (α if *X*) (β if –*X*). Then:

[17] Bradley borrows this reading of Ramsey's Archimedean axiom from Robert Koons (1993, p. 148).

$p(X) =_{\text{def}} [v(\zeta) - v(\beta)]/[v(\alpha) - v(\beta)]$. It also follows from the 11 Ramsey-axioms that for an arbitrary proposition X:

1. (i) $p(X) \geq 0$
   (ii) $p(X) + p(-X) = 1$
   (iii) $p(X|Y) + p(-X|Y) = 1$
2. $p(X|Y) = p(X\&Y)/p(Y)$

$p(\cdot)$ is therefore a probability function.

Let us define an action a as a function from a current state of the world (or rather, a proposition which corresponds to a current state of the world) X to an outcome α: $a: X \rightarrow \alpha$. Since the Ramsey-axioms allow us to define the valuation function $v(\cdot)$ on preferences and the probability function $p(\cdot)$ on propositions, these axioms suffice for defining the expected value of an action a:

$$E\left[v\left(a\right)\right] = \sum_{i=1}^{n} p\left(X_i\right) v\left[a\left(X_i\right)\right]$$

for each possible current state of the world X_i. We can then assert:

$$E\left[v\left(a\right)\right] \geq E\left[v\left(b\right)\right] \leftrightarrow \sum_{i=1}^{n} p\left(X_i\right) v\left[a\left(X_i\right)\right] \geq \sum_{i=1}^{n} p\left(X_i\right) v\left[b\left(X_i\right)\right]$$

Suppose, moreover, that we assume the following principle of expected value:

(EV): A rational agent will weakly prefer an action a to an action b iff the expected value of a is no less than the expected value of b.

Then we will be able to assert the following decision rule:

$$a \geq b \leftrightarrow E\left[v\left(a\right)\right] \geq E\left[v\left(b\right)\right] \leftrightarrow \sum_{i=1}^{n} p\left(X_i\right) v\left[a\left(X_i\right)\right] \geq \sum_{i=1}^{n} p\left(X_i\right) v\left[b\left(X_i\right)\right]$$

An agent with rational preferences over outcomes will thus have a preference-order over actions as well. That preference-order over actions, moreover, will be rational, in that it will satisfy a set of decision-theoretic axioms.

The agent who judges that a is preferable to b (and all his other options) will normally decide to perform a, and thus will intend to do so. And, when the time comes, he normally will do so. In using this model to represent deliberation, decision, intention and action, there is no implication that all conscious deliberation proceeds through the construction of quantitative probability judgments and valuations. Nor that all action, or even all intentional action, is preceded by conscious deliberation. Nor that all action, or even all intentional action, is preceded by a conscious decision about what to do, whether preceded by deliberation or not. Nor even that all action is intentional. Decision theory is an *interpretive* framework. It gives us a mathematically useful and convenient way to *represent* the agent who is

observed to act in a way consistent with a rational set of preferences over actions *as* having such preferences, and *as* having arrived at those preferences by way of (an admittedly idealized form of) rational deliberation. It thus provides a model of rational deliberation leading through judgment and decision to action. We need not assume that such an agent actually acted on the basis of deliberation over expected values, or even on the basis of any conscious deliberation. If we take a Bayesian view of the brain, decision theory also provides us with a way of modeling the valuations and probability calculations that take place, encoded at the neural level, prior to all voluntary movement, with or without the agent realizing that they are taking place.[18] But we can make use of decision-theoretic representations without committing ourselves to a Bayesian theory of the brain—though I myself am inclined to accept such a theory.

Decision theory—or rather, those versions of it which assume agents have preferences over outcomes rather than actions—does provide us with a deliberative method by which we may select the best means to our end, given sufficient information. Let us take one of Richard Jeffrey's simple examples: you have been invited to dinner, you have one bottle of red wine and one of white, you can only bring one of them (perhaps your roommate has requested one of either color be left for him), and you are uncertain whether beef or chicken is being prepared (though you are certain it is one of the two) (Jeffrey 1990, p. 74). Suppose you are also confident, though not certain, that the cooking will begin after you arrive, and what is prepared will be selected to match the wine. Assume your preferences are [chicken & white ~ beef & red > chicken & red > beef & white]. That is, with respect to the end of eating a delicious meal (which serves as the covering value for your preference judgment), you judge the constitutive means of eating a meal of chicken with white wine or beef with red wine to be equally good, and each better than chicken with red, which is in turn better than beef with white. Assign values representing these preferences:

	Chicken	**Beef**
White	1	-1
Red	0	1

Now represent the probabilities you assign to the various possible outcomes:

	Chicken	**Beef**
White	.75	.25
Red	.25	.75

[18] For some examples of the Bayesian current within contemporary neuro- and cognitive science—according to which actual human cognitive processes, at the neural level, can be modeled quite accurately using Bayesian principles and are often executed in a manner which is fairly close to what would be considered optimal within a Bayesian model—see (Oakesford and Chater 1999; Doya et al. 2007).

The probabilities reflect your confidence that if you bring white, it is more likely than not that chicken will then be prepared, (and likewise for red). Your goal is to eat either a meal of chicken with white or beef with red. These experiences rank highest of the ones available to you, insofar as they are realizations of the more general end of eating a good meal this evening. Bringing a bottle of wine that will complement the dish is the instrumental means to this end over which you can exercise control. The question is which means to select—bringing red or bringing white—given all this information, and our framework allows us to work out an answer. We can calculate the expected value of bringing white wine at 0.5 and that of bringing red wine at 0.75, based on the above preferences and probabilities. So bringing a bottle of red wine is a better means to your end than bringing a bottle of white. You have a better shot at getting the delicious meal you are aiming at if you bring red. Bayesian decision theory, then, does provide us with a very limited model of deliberating about selecting means. Given (a) a list of the potential outcomes of one's available actions; (b) a preference ranking over those outcomes; and (c) a probability distribution relating each means to each outcome, decision theory tells us how to determine which available means is best.

With respect to action-directed practical reasoning, Bayesianism dramatically shifts our perspective on decision-making from the perspective of the "standard model". As we will see, appreciating the nature of this shift is an important step toward developing a model of ends-deliberation. There is no special work to be done in the Bayesian model by the notions of necessary or sufficient means. We are now concerned only with the best available means, understood as the one with the highest expected value. Deductive validity is now off the table even as an ideal to which practical reasoning should try to approach. Audi's deductive model for practical reasoning has been replaced by a probabilistic one, which is akin to his notion of reasonable practical inference, though there are important differences. Audi's division of reasonable inference into the minimally adequate, the standardly adequate, and the cogent can be left behind, along with the notion that practical reasoning is itself a form of "inference" traditionally conceived. Bayesian deliberation yields an expected value for each available action, and then recommends the action with the highest such value as the best available means, and so as the action that the agent should perform. The conclusion that ϕ is the action the agent should perform is no longer derived from a set of premises in the way Audi imagines it to be. That the agent should ϕ is now merely another way of stating that ϕ-ing has the highest expected value. If there is any sort of action-directed practical inference here, it is nothing like Audi's, but rather the following rather uninteresting one:

1. I should do that which has the highest expected value.
2. ϕ-ing is that which has the highest expected value.
3. Therefore, I should ϕ.

This inference is valid insofar as it is nothing more than a straightforward substitution. Since the inference is valid, it does not make sense to ask whether the conclusion is more likely to be true than false given the premises; so we do not need Audi's grades of reasonableness. But notice how different the first premise is from

the one in Audi's inference. Rather than stating a want, the first premise here already makes an assertion about what action should be done—the fairly innocuous assertion that one should do that which is the best means to one's end. Which particular end one has settled on is relevant to the determination of which action has the highest expected value. We see the influence of the chosen end in the fact that ϕ-ing, and not some other action, appears in the second premise. ϕ-ing appears there for the reason that it is the best means to the particular end chosen by the agent. But all the deliberative work is done behind the scenes in determining which action to put in the second premise—in determining that it is ϕ-ing that is the best means. It is in making that determination that the agent uses his preferences over potential outcomes, the "wants" which Audi places within the inference to action itself. Insofar as this chain of reasoning is a probabilistic one, it bears some resemblance to Audi's "reasonable inferences"; but Bayesian reasoning terminates in the selection of a means, and its structure does not even approximate that of a valid deductive inference. The Bayesian model is thus *primarily* a model of deliberation about means-selection. Rather than being an alternative to or a more precise version of the standard model, it is an attempt to capture a different process: the process of determining which means is best given one's end.

We must be careful about how we interpret the principle of expected value and the leftmost biconditional above. There is no mathematically necessary or provable connection between the expected value of the agent's available actions and his preferences over those actions, given the version of decision theory we have employed here. This is due to the fact that we are not working with a version of decision theory that assumes the agent has preferences over actions to begin with, and then specifies axioms that that preference set must satisfy such that an expected value representation of those preferences can be given. We have instead assumed that the agent has preferences over outcomes, and specified the axioms that restrict that preference set. The principle of expected value is assumed as a bridge between the agent's preferences over outcomes and probability judgments over states of affairs, on the one hand, and his preferences over actions, on the other. For our purposes, moreover, we are to interpret this principle as an additional requirement of rationality, though of a different type from the axioms given above. The principle is a requirement of rationality in the sense of being a requirement of practical reasoning—a requirement on the move from preferences to choice. The axioms pertain to the rationality of sets of preferences, by restricting the structure that such sets may have. Insofar as the agent is rational in the practical reasoning-sense, he will prefer one action to another if, and only if, that action has the higher expected value.

We might add that, insofar as he *ought* (on the balance of reasons) to act rationally, the agent *ought* to prefer the option with the highest expected value. This is a *normative* interpretation of the principle. There has been much debate in recent years over whether rational requirements (in the sense of requirements of practical reasoning) are normative—over whether we ought, on the balance of reasons, to adhere to them. Many of the arguments that rational requirements are not normative requirements proceed from the assumption that rational requirements have a narrow-scope. Take the rational requirement that one ought to do what is necessary to

achieve one's end. If this has narrow-scope, then if I intend to ϕ, and I must ψ in order to do so, then I ought to ψ, *whatever ϕ and ψ are.* I reject such arguments on the grounds that this assumption is false: rational requirements are wide-scope requirements. To say that it is a rational requirement that one do what is necessary to achieve one's end is to say that if I intend to ϕ, and I must ψ in order to do so, then rationality requires that I *not* continue to intend to ϕ, and yet not do ψ.[19] I believe that what rationality *does* require that I do is either change my end/intention, or do ψ, *depending on what I have most reason to do*. This is a requirement of *rationality* in virtue of the fact that if what I have most reason to do is change my intention, there will be reasons to believe that this is so, and rationality requires that I search for and respond correctly to these reasons. The rational revision of one's ends and intentions is the subject of the next chapter, where the theory of ends-deliberation being built up to here will be introduced. To be rational, then, is to seek out and respond correctly to reasons, in one's judgments, deliberations, intentions and actions.[20] And this is of course bound to be precisely what I ought to do, in the normative sense. So I believe rational and normative requirements are *co-extensive*. Of course, if doing so is not feasible, then I cannot be rationally required to do so—but then neither, I think, can it be said that I ought to change my intention, *in any interesting and practical sense of the normative 'ought'*. I will have more to say on this point in Chap. 13.

The other argument that rational requirements are not normative—that, at least some of the time, we ought not be rational—appeals to what are known as "state-given reasons." These are alleged to be reasons for being in some state—such as a state of intending to do something—which are grounded by the fact that some benefit is to be had in return for being in the state itself. So for example, if one will receive some benefit for intending to do something noxious, but one need not actually perform the noxious action to receive the benefit, one is supposed to have a state-given reason for intending to do it—a reason which has nothing to do with the object of the intention, i.e. the action one intends to do or any outcome of that action. One then allegedly has a reason to intend to do something which one has no reason to do. If intending to do what one has no reason to do is irrational, one thereby has a reason to be irrational. And if the benefit is great enough, then being irrational may be what the agent ought to do. So, the argument goes, there are cases where we ought to be irrational. The *locus classicus* for such scenarios is of course Gregory Kavka's "toxin puzzle" (Kavka 1983).

This is by no means the appropriate time to become entangled in this complex debate. I will note, however, that I do not believe there are any such things as state-given reasons, and I am uniformly unpersuaded by arguments that attempt to establish their existence. All toxin-puzzle-style arguments that there are state-given reasons for intending to pursue what is noxious rely on the same hidden, and profoundly mistaken, assumption: that intending to ϕ is a mental action which is a

[19] For vigorous replies to objections to a wide-scope interpretation, see (Rippon 2011).

[20] As Raz puts it, "[R]ationality is the ability to realize the normative significance of the normative features of the world, and the ability to respond accordingly" (Raz 1999a, p. 68).

means to the end of ϕ-ing—where ϕ-ing is an action that involves bodily movement. This assumption makes intending to ϕ out to be a kind of intermediate action which an agent performs in order to cause himself to ϕ. The whole point of these arguments, whether their authors realize it or not, is to construct scenarios in which the mental action of intending to ϕ, in the absence of any reason to ϕ, becomes a means to something else, in that it causes some other agent to act to bring about some other end, which the intending agent does have a reason to achieve. *But intending is not a means to anything*. It is *not even* a means to the performance of the action intended. As GEM Anscombe has argued, it is *not* the case that "the relation of *being done in execution of a certain intention*, or *being done intentionally*, is a causal relation between act and intention" (Anscombe 1983/2005, p. 95). Rather, the relation between intention, in all its guises, and action is teleological.[21] Following Anscombe, we may distinguish three sense of "intending to ϕ," or "having an intention to ϕ" (Anscome 1963, p. 1). One sense is "acting *with the intention* of ϕ-ing." To say that an agent acts with the intention of ϕ-ing is just to say that the agent acts, and that ϕ-ing itself, or some expected outcome of ϕ-ing, is the/an end of the agent's action. Another sense of "intending to ϕ" is "having a *prospective intention* to ϕ." To say that an agent has a prospective intention to ϕ, or an intention to ϕ in the future, is to say nothing more than that ϕ-ing, or some expected outcome of ϕ-ing, is one of the agent's ends, and the agent has decided, or settled on a plan, or "made up his mind," to act with the intention of ϕ-ing at some future time (perhaps some specific time in the future, perhaps just some time or other in the future.) A final sense of "intending to ϕ" is "ϕ-ing *intentionally*." But to say that an agent ϕ's intentionally is just to say that the agent ϕ's with the intention of ϕ-ing. In none of these cases is intending a means to acting. Indeed, even in the sense in which one intends to ϕ *mere moments* before one ϕ's intentionally, intending *cannot* be a means to acting. As is well known by now, the motor cortex reaches its readiness potential for

[21] I take this to be true even in the case of purely expressive intentional actions. For a defense of that claim, see (Raz 1999b). I must emphasize that my remarks here are certainly not meant to be taken as anything approaching a complete theory of intention. My goal is simply to say enough about the view of intention I accept to show why I am unconvinced by arguments for the existence of state-given reasons. There are many important questions that any complete theory of intention would have to answer which I cannot consider here, and many intricate debates surrounding attempts to answer those questions which I cannot discuss. For example, I have certainly not given a 'unified' account of intention, since I do not explain prospective intentions in terms of intention in action—only in terms of decision to act with an intention in the future. And I have said nothing about whether one must know, or believe, that one is ϕ-ing in order to ϕ intentionally. I am not sure whether I think this is the case—though I certainly think that one cannot act intentionally without believing that one is acting in some way or other. I will note that two of the most pressing issues for a teleological view of intention–the question of how intention relates to judgment about the good, and how intention relates to weak-willed actions—are issues I will confront over the next four chapters. A frequent concern about the sorts of claims I make here is that they render intention epiphenomenal and/or eliminable. For more on this point, see note 8 in Chap. 13.

voluntary movement *before* an agent becomes consciously aware of intending to move (Kornhuber and Deecke 1965).[22]

What toxin-puzzle cases show, as others have pointed out, is that an agent can have a reason to cause himself to intend to do something (by going to a hypnotist, say) even though he has no reason to intend to do it. The strongest objection that has been made to this response is that, since intending to ϕ follows analytically from causing oneself to intend to ϕ, the response entails that one can have a reason to do something, but not a reason to do what follows analytically from it (Reisner 2011, pp. 44–45). But once we see that intending is not the sort of thing that is a means to anything (unlike the action of causing-oneself-to-intend something, which may be), we can see that the sort of reason one might have for causing-oneself-to-intend—viz., an instrumental reason, a reason to act for the sake of achieving some end—is very different from the sort of reason one can have for intending. Since the former may be a means to any valuable end (as the toxin puzzle shows), that end may provide a reason to do it. But since intending is not a means, the only reason one can have for intending to do something is the reason there is for doing it—indeed, the only sense in which one can have a reason for intending is the sense in which one can have a reason for doing what is intended.[23] So we should not be surprised by the fact that one can have a reason to cause-oneself-to-intend while lacking a reason to intend, despite the analytic relationship between the two.

Though I am unconvinced by the arguments that purport to show that rational requirements are not normative, I am happy to acknowledge that the rational and the normative are nonetheless conceptually distinct. The latter concerns what there is (most) reason to do or believe, the former concerns the ways in which we respond to what reasons there are with respect to our judgments, deliberations, intentions and actions. Again, I will have much more to say about the nature of reasons and the ways in which we respond to them in Chap. 13.

Let us return, then, to more immediately pressing matters. Most decision-theoretic models developed since Ramsey have no need to introduce the principle of expected value as an independent assumption in order to induce a preference-ranking over actions. This is precisely because they begin by assuming that the agent has preferences over actions (rather than over outcomes), and introduce axioms which restrict those preferences over actions. The aim of most contemporary decision-theoretic models is to show that if an agent's preferences over actions satisfy the axioms, it will be as if the agent has ranked his possible actions according to their expected value. What is most desired in a decision theory of this type is to obtain just such an expected value result from a set of axioms whose members can all be independently defended as constitutive of rational preference. It would then follow, purely as a consequence of an independently defensible account of rationality, that rational agents can be treated as expected value maximizers with respect to

[22] The work of Kornhuber and Deeke has spawned a large literature on Bereitschaftspotential (Readiness Potential). For a recent, up-to-date survey, and a philosophically sophisticated discussion of the implications of research into this topic, see (Kornhuber and Deeke 2012).

[23] For a related response to the toxin puzzle from a Sellarsian perspective, see (Heath 2008, ch. 5).

their choices about how to act. There is as yet no version of decision theory that achieves this goal: all decision-theoretic models require purely structural axioms which have nothing to do with rationality, and are incorporated solely because they are required in order to obtain the expected value result. This is as true of a Ramsey-style theory as it is of any other. But many authors have seen it as a point against Ramsey's theory that it must, in addition, introduce the principle of expected value as an independent assumption in order to obtain a preference-ranking over actions. So why have I decided to use Ramsey's theory as the foundation of my dynamic model of ends-deliberation?

We will see below that a Ramsey-style theory like the one just introduced has a number of strengths missing in other theories, and lacks a number of their weaknesses. But it is prudent to address this apparent weakness of a Ramsey-style theory up front. The answer, in short, is that from the perspective of my project, this feature of the theory merely appears to be a weakness, and is in fact a strength. The source of this affinity between my project and Ramsey's theory lies in the fact that I aim to produce a *dynamic* account of a rational agent's deliberations about what ends to pursue. The actions which are preferred by an agent faced with a series of decision about how to act should depend on what that agent hopes to accomplish. When two agents have different preferences over actions, this difference must be explained by appeal to what the agents believe about the results their available actions will lead to, and how they view the value of those results. The agent in my model is engaged in the process of determining precisely what value he places on the various ends he might be able to achieve. This is a problem which must be resolved prior to his formation of any preferences over actions. He cannot have a rational preference for what to do without first having preferences over what is likely to come of what he might do. This is true not only of those actions that the agent will come to see as having merely instrumental value, but of those actions which he will come to see as having final or intrinsic value as well. In order for an agent to prefer to engage in some activity for its own sake, he must view that engagement as valuable in itself. It is just such a conclusion—about what is of value and what is not—that the agent in my model is in the process of formulating. My model, therefore, requires a decision-theoretic foundation that does not assume that agents already have preferences over actions, since such preferences are something that the agents in my model are only beginning to work toward through the prior process of deliberating about their preferences over ends.[24]

[24] Richard Bradley argues, quite correctly in my view, that any criticism of Ramsey's theory on the grounds that it must independently assume the principle of expected value is misplaced. Ramsey's project was not to find a set of rationally justifiable axioms on preference which would themselves serve as a justification for the principle of expected value as a theoretical claim about rational belief and desire (once that principle could be shown to follow from those axioms). His was the quite different project of formulating an elegant solution to the problem of defining and measuring the variables of decision-theory—degrees of partial belief and desire. I agree with Bradley that Ramsey's work, considered from this vantage point, remains unsurpassed (Bradley 2001).

4.3 Desirability, Evidence, and Causation

Ramsey's model allows us to represent numerically the preferences of a rational agent over outcomes, which we have defined as near-maximal propositions about possible worlds. But the type of decision theory I require as the foundation of my account must be capable of representing an agent's preferences over any set of arbitrary propositions. The valuation assigned to a proposition is generally called the *desirability* of that proposition. So we might ask, from the perspective of a given agent, how desirable it is that it rain all afternoon tomorrow in the city where he lives, without having to fill that proposition out with enough other (irrelevant) details to make it a near-maximal world-proposition. Once we shift our focus from the value of outcomes to the desirability of propositions more generally, we must introduce the distinction between *evidential* decision theory and *causal* decision theory.

To see how and why this distinction emerges, let us consider a different way of representing actions within a decision-theoretic model. We previously viewed actions as functions from states to outcomes—with both states and outcomes represented in the model as propositions which describe those states and outcomes. But we can represent actions themselves as propositions within the model as well—and thus gain the advantage of working with a theory which requires only propositions in its ontology. Each possible action will then be identified with the proposition that the agent performs that action. More precisely, an "action" is a proposition which describes an event whose occurrence an agent has complete control over—that is, an event which an agent can bring about at will.[25] We can then ask how desirable, for a given agent, is the truth of the proposition that that agent perform that action. It is when we ask this question that the distinction between causal and evidential decision theory emerges. The desirability of a proposition is a function of the values of the various total possible states of affairs with which the truth of that proposition is consistent:

$$des(X) = \sum_{\alpha \in \Gamma} p(\alpha \, given X) \cdot v(\alpha).$$

If the proposition in question is an action-proposition, the expression "the probability of outcome α given the truth of proposition *X*" can be interpreted in two different ways. On one interpretation, the "given" is taken to signify the familiar notion of conditional probability. The expression is then read as "$p(\alpha|X)$" and denotes the probability that α is true on the evidence of the (assumed) truth of *X*. Under this interpretation, what the above expression gives is the evidential desirability of the truth of *X*. This notion of evidential desirability is the foundation of evidential decision theory. The evidential desirability of a proposition *X* is often referred to as the "news value" of the proposition—a measure of how welcome the

[25] For the importance of this rather narrow understanding of actions in a decision-theoretic context, see (Joyce 1998, p. 61).

news of the proposition's truth would be to the agent. On the other interpretation, the "given" is taken to signify counter-factual dependence. The expression is then read as "$p(\alpha \rightarrow X)$" and denotes the probability that α would be true if X were true. A high probability signifies that the action referred to in X is highly causally efficacious in bringing about the outcome referred to in α. Under this interpretation, what the above expression gives is the causal desirability of the truth of X. This notion of causal desirability is the foundation of causal decision theory. The causal desirability of X is often referred to as the "efficacy value" of the proposition.

The distinction between evidential and causal decision theory is not a merely conceptual one; the evidential and causal desirability of action-propositions can come apart. This occurs in decision problems in which the fact that an action is performed is good evidence that some outcome will occur, even though the action has no causal power to bring about that action. Such cases are called Newcombe problems, after William Newcombe who first formulated them. There is a large and interesting literature on these problems, but the details need not concern us here.[26] What does matter for our purposes is the fact that these two forms of decision theory must be distinguished, and that—as we will see—the resources of both will be required for the full development of my account. The question we are faced with, therefore, is whether the Ramsey-style decision theory we have begun with as our foundation can be extended to evidential and causal desirabilities. Fortunately, David Sobel has answered this question in the affirmative (Sobel 1998). Once we have valuations assigned to outcomes, we can define the causal and evidential desirability of an arbitrary proposition in precisely the natural and intuitive way just described. This lets us see another of the advantages of a Ramsey-style theory: it provides a neutral foundation for both evidential and causal decision theory.

With respect to causal decision theory, my project will only require that we be able to define the causal desirabilities of action-propositions. The Ramsey-style theory which serves as my foundation is, as we have just noted, adequate to this task. But with respect to evidential decision theory, which will provide the formal language in which my dynamic account of ends-deliberation is expressed, my project requires a fully developed representation theorem. Once again, the Ramsey-style theory suffices as a foundation. Richard Bradley has proved that with a few modifications, the set of axioms introduced above can be used to derive an evidential decision theory which possesses many advantages (Bradley 2004). It is this theory that I will apply in formulating my dynamic account.

Bradley calls a proposition P ethically neutral iff it is neutral with respect to all propositions Q with which it is consistent. P is neutral with respect to consistent proposition Q iff $P\&Q \sim Q \sim P\&\text{-}Q$. We can then obtain a Ramsey-style evidential decision theory from axioms R1 – R6, R8 – R10, plus an axiom which replaces R11 and which Bradley calls the principle of ethical conditionalism:

[26] See (Joyce 1998, ch. 1), for a vivid introduction to Newcombe problems; and Joyce's bibliography for many of the most important works in the Newcombe literature.

(EC): For any propositions P and Q there exist propositions P' and Q' such that P' implies P, Q' implies Q, and $P' \sim Q'$.

The outcome-propositions in R1-R10 are replaced with prospect-propositions which need not be near-maximal. Dispensing with near-maximal world-propositions makes R7 redundant (Bradley 2004, p. 491).

We can then speak in terms of the expected value of an arbitrary proposition X's being true, as the product of the desirability of the proposition's truth and the agent's degree of credence in its truth: $p(X)des(X)$. So if X is "It rains in New York City on May 1st 2017," then X is true if, and only if, it rains in New York City on May 1st 2017; $p(X)$ is the agent's judgment of the probability that this is true; $des(X)$ is the desirability to the agent of its being true; and the product of these two is the expected value to the agent of its being true (which is not to be confused with the expected value of a wager that it is true). The desirability of X will be a probability-weighted sum of the ways X might be true—the possible states of affairs which are consistent with the truth of X:

$$des(X) = \sum_{E} p(E \mid X) \cdot v(X \,\&\, E)$$

for some partition of state-propositions $E = \{E_1, E_2, \ldots, E_n\}$. We can define the probability that some proposition X is true given the performance of some action (or rather, the truth of some action-proposition) A as $p(X|A)$, and the expected value of the truth of X given the performance of A as $V(X) = p(X|A)des(A\&X)$.

We can also compute the desirability and expected value of action-propositions themselves, even though the theory does not assume that the agent has preferences over actions. Suppose the set of propositions $\{X_1, \ldots, X_n\}$ is the set of propositions describing events which are possible outcomes of some action A. Then the desirability of A is the sum of the desirabilities of these outcome-propositions—these are the ways the world might be, consistent with the performance of the action—weighted by the probability of each outcome-proposition being true, given the performance of the action:

$$des(A) = \Sigma_X \, p(X \mid A) des(A \,\&\, X).$$

The expected value of the action A, $V(A)$, is then:

$$V(A) = p(A \mid A) des(A),$$

And since $p(A|A) = 1$, we have

$$V(A) = des(A) = \Sigma_X \, p(X \mid A) des(A \,\&\, X).$$

The expected value of an action, then, is equal to the sum of the expected values of its possible outcomes, as we have defined these values. The agent's degree of credence in the truth of X will change according to the evidence available to the agent

that bears on the truth of X. For a partition of propositions $E = \{E_1, E_2, \ldots, E_n\}$ which bear on the truth of X:

$$p_{new}(X) = \sum_{i=1}^{n} p(X|E_i) \cdot p_{new}(E_i).$$

This completes the development of the decision-theoretic model which provides the required background for my dynamic account of ends-deliberation. Before proceeding to examine some recent developments in decision theory which influence the shape my account will take, I must address the issue of the alleged weaknesses of the Ramsey-style theory I have chosen as my foundation, and lay out some of the strengths of this foundation which make it uniquely appropriate to my project.

4.4 Advantages and Alleged Disadvantages of a Ramsey-Style Decision Theory

We have already taken notice of a couple of the important advantages of a Ramsey-style decision theory. The theory provides an elegant and intuitively appealing way to solve the problem of measuring the variables of decision-theory, degrees of partial belief and desire. And it provides a neutral foundation for defining both evidential and causal desirabilities, as well as having a natural extension to a fully developed evidential decision theory. The theory is also especially well suited to the role which I need a decision theory to play: that of background theory for a dynamic account of rational deliberation about ends. We have noted that one feature that makes it so is the fact that it defines preferences over outcomes, rather than actions, and takes on the principle of expected value as an independent assumption. These are the same features which allow it to serve as a model of the process of instrumental deliberation, as we saw in this chapter. We will see below that another such feature is the fact that it is values of outcomes—which is to say, sets of indifferent outcomes—which the theory takes as the constituents of an agent's preference-ranking. This fact will prove important in developing a way to model deliberation about incorporating newly discovered potential ends into the agent's preference-ranking.

The Ramsey-style evidential decision theory shares two important virtues of Bolker-Jeffrey decision theory, which is one of the leading versions of evidential decision theory. It is atomless: every proposition in the domain of preference can be defined in terms of other propositions, and thus the theory contains no atomic propositions. This is required by axiom EC (Bradley 2001, p. 21). And it is partition invariant: the evidential desirability of a proposition X can be defined with respect to any partition of state-propositions consistent with X—any set of mutually exhaustive descriptions of the ways X might be true. Partition invariance is one of the great benefits of interpreting the desirability of a proposition as its news-value (Joyce

1998, p. 122). It is a feature which allows a decision theory to be applied in so-called "small-world" decision problems: problems in which the possible outcomes are differentiated coarsely, without specifying every miniscule variation that might conceivably be of interest to the decision-maker. These are, of course, the sorts of decision problems we actually face, as it is always practically impossible to specify the potential outcomes of our available actions down to the last detail. But a Ramsey-style evidential decision theory also has a great virtue which Bolker-Jeffrey decision theory lacks. The axioms of Bolker-Jeffrey theory cannot be used to derive a uniqueness result—the valuations that occur in that theory are not unique up to a positive affine transformation.[27] As we have seen, the axioms of a Ramsey-style theory are sufficient to secure a uniqueness result.

A Ramsey-style theory also has a number of important advantages over the decision-theory of L. J. Savage. Bradley has claimed that Ramsey's axioms lack the independent justification of Savage's, and that this is due to the fact that Ramsey was concerned with articulating a theory of measurement for decision theory and so did not shy away from incorporating a number of purely structural axioms, whereas Savage was concerned with articulating a theory of rationality, and carefully chose axioms which were, individually, plausible candidates for restrictions on rational preference (Bradley 2004, pp. 493–494). But this claim ignores the fact that Savage's theory incorporates a number of extremely implausible requirements, and that neither it nor any other decision theory in existence is free of purely structural axioms which lack independent justification as canons of rationality. As James Joyce has forcefully argued, Savage's requirement of the existence of constant acts is one of the most implausible features of any decision theory (Joyce 1998, pp. 65–67). A constant act is an act that always produces the same outcome, regardless of the state in which it is performed. Savage requires a constant act for every possible outcome. Thus, if it is possible for me to be deliriously happy, then there must be some act which I could, at least in principle, perform, that would result in my being deliriously happy, even in a circumstance in which I was faced with the imminent destruction of the world, and the knowledge that my own life and the life of everyone I love was about to come to a hideous end.

Another implausible feature of Savage's theory, which Bradley himself notes, is the requirement of state-independence (Bradley 2004, p. 494). Savage assumes that the value of the outcome of any action is completely independent of the state of the world in which that outcome is realized. But as Bradley observes, "it is a perfectly banal fact about our attitudes to many things that they depend on all sorts of contextual factors" (Bradley 2004, p. 494). Let us say that the outcome of my action of drinking hot chocolate is that hot chocolate is consumed by me. I may find the consumption of hot chocolate to be delightful on a winter's night but not so on a summer's day. The value to me of this outcome thus depends on the weather. Savage's theory is forced to deny this perfectly banal fact. It must assume that whatever the value to me of consuming hot chocolate is, it must be constant from January to July. In a Ramsey-style theory, on the contrary, the desirability of any outcome that is not

[27] For the weaker Bolker-Jeffrey uniqueness theorem, see (Jeffrey 1990, ch. 8).

ethically neutral depends on the state in which it is realized; outcome desirabilities are derived from the desirabilities of the ways in which those outcomes might be realized.

One final advantage of Ramsey-style decision theory, at least in the evidential version developed by Bradley, is that it is consistent with holism in the theory of value. The essential mark of a value-holist, according to Jonathan Dancy, is a commitment to the claim that whatever has value in some context may have a different value, an opposite value, or no value at all, in another context (Dancy 2006, p. 331). This is precisely the idea that lies behind the axiom of ethical conditionalism in Bradley's Ramsey-style evidential decision theory. Bradley's own interpretation of the axiom is that "however good (or bad) some possibility might be on average, there are imaginable circumstances in which it is not so. No prospect is good or bad in itself, but is only so relative to the conditions under which it is expected to be realized" (Bradley 2004, p. 490).

Dancy is committed not only to value-holism, but also to the further doctrine of choice-holism, which holds that the value of one available alternative can be affected by the nature of the other available alternatives (Dancy 2006, p. 339). It seems to me that due to its rejection of the requirement of state-independence, a Ramsey-style theory is consistent with this view as well. Dancy is concerned with examples of the following sort:

> Suppose that I have to buy a house in Reading, and have a choice between a smaller house within walking distance of the university, and a larger and more expensive house that requires a bus ride. I prefer the larger one despite the bus ride. Then a third house, even larger but also farther away than the second, comes onto the market. I realize that if I buy the second house, I will always regret not having bought the third. With this in mind, I buy the first house. Is this rational? I suggest that it can be. (Dancy 2006, p. 336)

Part of the state in which a prospect A might be realized is the fact that (a chance at) prospect B was passed over in favor of A. Just as my preference for hot chocolate or lemonade can depend on the weather, my preference for one prospect or another can depend on what other prospects are available. My obtaining B without regret is a different outcome from my obtaining B with regret; and the state in which B is realized without C having been a possibility is different from the state in which B is realized with C having been mine for the taking. Since the valuation of outcomes depends on the states in which those outcomes are realized, the outcome of owning the second house may receive a different valuation in one of these states than it receives in the other.

A Ramsey-style decision theory certainly does require that the agent have a complete preference order, and Dancy believes that choice-holism and completeness are incompatible. His main argument for this conclusion is as follows:

> Suppose that we have a full ordering of all relativized options…Suppose now that I ask of item 32 in the list how it compares in value with item 33. It need not be the case that my answer is that item 32 is more valuable than item 33. The option '33 when I could have had 32' is a different option from the simple option '33', no matter how internally complex option 33 may be—and the same goes for option 32. But if my ranking order does not even

> commit me to claims about the relative values of the items ranked, it is pointless. (Dancy 2006, pp. 342–343)

To see what is wrong with this argument, let us consider another example of the type Dancy is concerned about. Call 'item 32' taking a 1 week beach vacation instead of going to work as usual. Call 'item 33' going to work as usual, despite having been offered a choice of a beach vacation or a mountain climbing vacation. In the latter choice situation, I choose to stay home because I am worried that if I were to choose the beach vacation, everyone would think me too timid to go mountain climbing—which I am, but I do not wish others to know this. I would rather be thought a workaholic, and miss out on the vacation. Dancy's concern is that even though the outcome "beach vacation instead of work" appears in my preference-ranking above the outcome "work instead of beach vacation or mountain-climbing vacation," I still cannot claim to value the former more than the latter. What I have not done is consider the *choice between choice situations*: the choice between (a) being offered a choice between work and a beach vacation and (b) being offered a choice between work, a beach vacation, and a mountain-climbing vacation. Dancy's point is that, given choice-holism, although I might prefer to find myself simply facing choice situation (a) and then choose the beach vacation, rather than find myself facing choice situation (b) and then choose to go to work—and thus I rank "beach instead of work" above "work instead of beach or mountain"—I still might prefer to choose to face situation (b) rather than face situation (a), supposing I must choose which choice situation to face, and am not simply confronted with one or the other.

Dancy seems to think that if (as choice-holism allows) I do prefer to choose to face (b) rather than to choose to face (a) when it is up to me which of these I face, then I cannot be said to value "beach instead of work"—the outcome I choose when facing (a)—more than "work instead of beach or mountain"—the outcome I choose when facing (b). For if I choose to face (b), then I will end up choosing "work instead of beach or mountain" *despite* the fact that I could have had "beach instead of work," if only I had chosen to face (a) instead. And yet *ex hypothesi* "beach instead of work" appears above "work instead of beach or mountain" in my preference-ranking. Choice-holism, then, seems to upset the claim that I have a complete preference-ranking, in which an outcome is valued according to how high in the ranking it appears. He thus concludes that choice-holism is incompatible with a complete preference-ranking.

But this conclusion rests on a mistake. The apparent inconsistency is merely apparent. In fact, the following three claims are perfectly consistent:

I. The agent most prefers a beach vacation to going to work, without having been given a prior choice about what set of outcomes to choose from (Dancy's 'item 32.')
II. The agent next prefers going to work instead of a beach vacation or a mountain climbing vacation, having been given no prior choice about what set of outcomes to choose from (Dancy's 'item 33.')

III. The agent least prefers going to work instead of a beach vacation or a mountain climbing vacation, having been given a prior choice of whether to choose an outcome from that set of three outcomes, or to choose an outcome from the set which includes only work and the beach (Call this 'item 34,' if you like.)

The problem is simply that the question Dancy wants to ask—"Would the agent choose item 32 if both 32 and 33 were available, or would he choose 33?"—is not a well-formed question. We can see that the outcomes which constitute items 32 and 33, when properly described, specify that the agent had no choice about which set of options he would be choosing from. That fact is part of the state of the world in which each of those outcomes is realized, just as much as facts about what other outcomes were available within a given menu. If it is possible for the agent to choose either "beach instead of work" or "work instead of beach or mountain," then he is *not* in a situation in which he is choosing between item 32 and item 33. For the outcomes that constitute items 32 and 33, when properly described, *exclude that very possibility*. What Dancy misses is that the outcomes in 32 and 33 are mutually inconsistent; they cannot be simultaneously available for the agent to choose between. But if Dancy's question is ill-formed, then what does it mean to say that the agent *prefers* item 32 to item 33? It means that the agent values the state in which he goes to the beach instead of work—having been simply confronted with that one choice, without having had to make any prior choice between menus of options—more than the state in which he goes to work rather than mountain climbing or to the beach—having likewise been simply confronted with that one choice. That preference is perfectly consistent with a preference for going to work rather than mountain climbing or to the beach, having first been given a choice between that menu of options and the reduced menu of working or going to the beach. And both of those preferences are perfectly consistent with a preference for *not having to make that first choice between menus*—a preference revealed by the location of what I called 'item 34' in the agent's preference-ranking. There is no inconsistency.

This argument, moreover, generalizes. It does not matter what example we use in place of Dancy's items 32 and 33, nor how simple or complex the structures of those items are. Part of an outcome, in a state-dependent theory, is the state in which the outcome (narrowly construed) is realized, and the state will always include the sorts of facts about the choice situation from which the outcome was chosen, which the above argument requires. If an agent has a complete preference-ranking of fully relativized outcomes, no questions about what the agent values are left unanswered, The truth of choice-holism does nothing to change that. A Ramsey-style evidential decision theory, then, is consistent with the basic tenets of both value-holism and choice-holism. I see this as an important virtue, since I regard both of those views as eminently plausible.

One genuine weakness in Ramsey's original model is a commitment to what John Howard Sobel calls "a thin logical atomism" (Sobel 1998, p. 236). Ramsey's ethically neutral propositions must be capable of being true or false independently of the truth or falsehood of any other proposition (Sobel 1998, p. 237). But since it

is not Ramsey's original model, but rather Bradley's atomless Ramsey-style evidential decision theory, which will provide the background for my dynamic model of ends-deliberation, my theory does not inherit Ramsey's commitment to even a thin logical atomism. I will make use of the definition of causal desirability which is derived directly from Ramsey's original model. But as we will see, causal desirabilities play a very small role in my account—they will be needed only in order to give a precise definition of the concept of an action which is a means to an end. My use of Ramsey's foundation for defining causal desirabilities, moreover, is based only on considerations of convenience and theoretical unity; I use it because the evidential decision theory my account relies on is based on Ramsey's original model. Those who find the presupposition of thin logical atomism in any part of the theory may be consoled by the fact that it is perfectly possible to define causal desirabilities without relying on Ramsey's original theory as a background, though doing so would require a more complex, pluralistic background for my own account.

The leading version of causal decision theory is that developed by Joyce. Joyce was the first decision theorist to prove a representation theorem for causal decision theory (Joyce 1998). In fact, he proved a representation theorem for a very general sort of decision theory which may be further specified as either a causal decision theory or an evidential decision theory. I do believe that it would be possible to use Joyce's versions of evidential and causal decision theory as my background, and thus avoid any of the objections that might be made to Ramsey's theory. My decision to use a Ramsey-style theory is based on the not inconsiderable theoretical virtue of simplicity. The development of Joyce's theory is quite complex, involving the analysis of indicative and subjunctive presuppositions, among other issues that have no bearing on the present task (Joyce 1998, ch. 7–8). The presupposition of thin logical atomism in the way I define means to ends seems a small price to pay, at least in the present context, for being able to use a background theory as straightforward and intuitive as Ramsey's.

5 Relevant Recent Developments in Decision Theory

There are a number of recent developments in decision theory which are relevant to the task of constructing a formal theory of ends-deliberation.

5.1 Meta-preference

Amartya Sen has made some suggestions regarding the form a model of ends-deliberation should take which, though they only amount to a bare sketch, are nonetheless a useful starting point for my enterprise (Sen 1982). Sen has argued that multiple preference-rankings, and orderings of those rankings themselves, are a key feature in representing the deliberation of autonomous agents (Sen 1982, pp. 80–83;

Sen 2002, pp. 615–618). An autonomous agent, he claims, must be one who is free to entertain different preferences-rankings, and free to revise his preferences on the basis of his own relevant considerations. Autonomous agents must be represented, then, as deliberating over possible preference-rankings and choosing between them. I am in full agreement with Sen that this is the key to representing the exercise of the capacity of autonomy. The next chapter is devoted to developing a precise account of this process.

5.2 Preference for Flexibility

David Kreps has successfully modeled a type of deliberation which resembles the sort of ends-deliberation that I am interested in (Kreps 1979). Kreps considers the case of an agent who is trying to choose between available menus (i.e. non-empty subsets of a choice set) (Kreps 1979). The choice of one item from within a menu cannot occur until some time in the future. The problem the agent faces is that he is aware that his preferences may be different at that future time from what they are now, but he is uncertain as to what exactly his preferences will be (Kreps 1979, p. 565). Kreps' goal is to model what he calls a "preference for flexibility": a preference for the menu of options which contains the options the agent is most likely to end up preferring. He introduces a set of states S which he identifies as the possible moods or tastes of the agent (Kreps 1979, p. 566). The agent is assumed to know the probability distribution over S. The agent's utility function is assumed to be state-dependent: the utility of a given item depends on the state the agent is in when he obtains that item. Kreps shows that it is possible to represent current preferences over menus in terms of the expected utility of choosing a given item from a given menu in a future state, where the probability of being in that state in the future is known.

Kreps' model is a more substantial move in the right direction as far as constructing a model of deliberation about ends. Kenneth Arrow has argued that something like Kreps' model can be used to model the capacity of autonomy in roughly the way I have conceived of it (Arrow 1995). The autonomous agent, on this view, would decide what preferences to have now on the basis of probability judgments regarding what his preferences will be in the future, so that the choices he makes now will move him toward outcomes which will be most preferable to him at the time that he achieves them. The agent is supposed to be autonomous insofar as he is using a rational decision procedure to determine what he should prefer. However, as Sen has argued, this is a poor way to represent the exercise of autonomy understood as the ability to reflect on what one should prefer and form preferences on the basis of that reflection (Sen 2002, pp. 619–620). The agent in Arrow's scenario does not determine for himself, on the basis of relevant considerations, what preferences he should have. Rather, he forms his current preferences based on his *predictions* of what acting on those preferences will eventually lead to, and on what he will happen to want in the future. This is why my own model will take preferences to be

judgments of choiceworthiness, and apply probabilities to those judgments being, from the position of the agent, correct. The autonomous agent will then adjust his preferences based on updating those probabilities in the light of relevant evidence, rather than forming his preferences based on the probability of finding himself in a given state.

5.3 *Reasons-Based Preference*

Franz Dietrich and Christian List have recently developed a model of "reason-based" preference and preference change (Dietrich and List 2013a).[28] The theory they develop is of great interest in its own right. But for our purposes, what is important is to understand the ways in which it fails to serve as a theory of ends-deliberation of the sort we are looking for. A brief examination of their theory will suffice to show this; it will, however, also help us map out the contours of the region of conceptual space into which a proper model of ends-deliberation must fit.

Dietrich and List take an important step toward providing a model of ends-deliberation, by modeling preference change endogenously—preferences for outcomes, in their theory, are based on reasons in favor of preferring those outcomes; and those reasons, and the relationship between those reasons and the outcomes they favor, are represented within the model. Dietrich and List, drawing on a long philosophical tradition, distinguish between *motivating reasons* and *normative reasons*. Motivating reasons are the features of potential outcomes that influence the actual preferences of an agent, and that actually motivate an agent to pursue (or avoid) those outcomes. They model the motivational value of an outcome x in terms of the motivating reasons that favor it, as weighted by a weighting function W (Dietrich and List 2013a, p. 26):

$$V_M(x) = W\left(\left\{R \in M : R \textit{ is true of } x\right\}\right)$$

An actual preference for an action A over and action B can then be represented in terms of the sum of the motivational values of the various potential outcomes weighted by the probabilities of attaining the potential outcomes of each action (Dietrich and List 2013a, p. 26):

$$A \geq_M B \leftrightarrow \sum_{x \in X} A(x) V_M(x) \geq \sum_{x \in X} B(x) V_M(x).$$

[28] There are many interesting similarities between Dietrich and List's mathematical model of reason-based preference, and the logical model of reason-based preference developed by Fenrong Liu (2011). Liu does an excellent job of formulating the task of developing a dynamic model of rational deliberation about basic preference. He finds, however, that within a purely qualitative-logical framework, no axiom set for such a model can be found.

Similarly, they model the normative value of an outcome x in terms of the normative reasons that favor it, likewise weighted by a function W:

$$V_N(x) = W\left(\left\{R \in M : R \textit{ is true of } x\right\}\right)$$

And a normative preference can then be similarly represented:

$$A \geq_N B \leftrightarrow \sum_{x \in X} A(x) V_N(x) \geq \sum_{x \in X} B(x) V_N(x).$$

A normative preference is a preference that an agent *should* have—a preference which is in fact supported by the balance of normative reasons.

Dietrich and List can then model actual preference-change in terms of changes in the agent's motivational set—the set M of reasons which successfully motivate the agent. Such changes may take the form of (i) reasons being deleted from the motivational set; (ii) reasons being added to the motivational set; or (iii) reasons in the motivational set being re-weighted. On the (entirely plausible) assumption that normative preference changes with changes in the agent's circumstances, normative preference-change can likewise be modeled in terms of changes in the normative preference set N. Finally, they suggest a way of modeling change in an agent's actual preference through deliberation on what (from a normative perspective) the agent has most reason to do. Suppose that R is a normative reason that applies to the agent, and that the agent believes that it is such. It is possible for the agent to *fix his attention* and *reflect* on R until R becomes *motivationally salient* for the agent. When it became motivationally salient it would enter the agent's motivational set M, and thus have an effect on the agent's actual preferences (Dietrich and List 2013a, p. 22).[29]

Modeling the changes in an agent's actual preferences which are brought about through deliberation on what the agent ought to do is precisely the goal of a theory of ends deliberation. But a genuine version of such a theory will provide an entirely endogenous way to model these changes. The short-coming of Dietrich and List's model, from this perspective, is that although it models preference-change endogenously, it models changes in the agent's motivational set exogenously. It takes changes in the content of an agent's motivational set as brute facts, and then models the effects of those changes on the agent's actual preferences. What their model leaves out is any representation of the process through by which an agent comes to believe that some fact or feature R is a normative reason that applies to him, and moves it into his motivational set (or that some R is not a normative reason, and moves it out of his motivational set); or by which an agent comes to re-weight a reason in his motivational set on the basis of having arrived at a new belief about the actual normative weight of that reason. All of Dietrich and List's talk of

[29] Dietrich and List have continued to explore this idea of preference-change being induced by shifts in attention which result in different features of options becoming motivationally salient (Dietrich and List 2013b).

attention-fixing, practical reflection, and belief about normative reasons is, in fact, purely interpretive; none of this is explicitly represented within the model.

A full-blooded theory of ends deliberation must provide an endogenous representation of changes in the agent's beliefs about what he ought to prefer. It ought to distinguish, moreover, between these beliefs and the agent's affective attachments, and explicitly represent both. And it ought to represent the distinct processes by which both beliefs and attachments change, and the way in which they come together to determine the agent' actual preferences. It should model exogenously only changes in the external world, in the agent's perceptions of the external world, and in the agent's attachments (when changes in the latter are not themselves caused by changes in the agent's beliefs about what he ought to prefer). The theory of Dietrich and List is capable of none of this. The theory I develop in the next chapter succeeds on every count.

References

Anderson, E. 1997. Practical reason and incommensurable goods. In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 90–109. Cambridge, MA: Harvard University Press.

Anscome, G.E.M. 1963. *Intention*, 2nd ed. Oxford: Oxford University Press.

Anscombe, G.E.M. 1983/2005. The causation of action. In *Human life, action and ethics*, ed. M. Geach, and L. Gormally, 89–108. Exeter: Imprint Academic.

Aristotle. *De Anima.* Oxford Classical Texts, ed. W.D. Ross. Oxford: Oxford University Press.

Arrow, K. 1951. *Social choice and individual values*. New Haven: Yale University Press.

Arrow, K. 1995. A note on freedom and flexibility. In *Choice, welfare and development: A festschrift in honor of Amartya K. Sen*, ed. K. Basu, P. Pattanaik, and K. Suzumura, 7–16. Oxford: Oxford University Press.

Audi, R. 2006. *Practical reasoning and ethical decision*. New York: Routledge.

Barberà, S., et al. 2004. *Handbook of utility theory*. New York: Springer.

Berlin, I. 1998. *The crooked timber of humanity*. Princeton: Princeton University Press.

Bradley, R. 2001. Ramsey and the measurement of belief. In *Foundations of Bayesianism*, ed. D. Corfield and J. Williamson. London: Kluwer.

Bradley, R. 2004. Ramsey's representation theorem. *Dialectica* 58(4): 483–498.

Bradley, R., and C. List. 2009. Desire-as-belief revisited. *Analysis* 69(1): 31–37.

Broome, J. 2006. Reasoning with preferences. In *Preference formation and well-being*, ed. S. Olsaretti, 183–208. Cambridge: Cambridge University Press.

Chang, R. 1997. Introduction. In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 1–34. Cambridge, MA: Harvard University Press.

Dancy, J. 2006. The Particularist's progress. In *Recent work on intrinsic value*, ed. Toni Rønnow-Rasmussen and Michael J. Zimmerman, 325–348. Dordrecht: Springer.

DeLancey, C. 2002. *Passionate engines*. New York: Oxford University Press.

Dietrich, F., and C. List. 2013a. A reason-based theory of rational choice. *Noûs* 47(1): 104–134.

Dietrich, F., and C. List. 2013b. Where do preferences come from? *International Journal of Game Theory* 42(3): 613–637.

Doya, K., S. Ishii, A. Pouget, and R.P.N. Rao (eds.). 2007. *Bayesian brain: Probabilistic approaches to neural coding*. Cambridge, MA: MIT Press.

Fischburn, P. 1972. *Mathematics of decision theory*. The Hague: Mouton.

Hausman, D.M. 2011. *Preference, value, choice and welfare*. Cambridge: Cambridge University Press.

Heath, J. 2008. *Following the rules: Practical reasoning and deontic constraint*. Oxford: Oxford University Press.
Jeffrey, R. 1990. *The logic of decision*. Chicago: University of Chicago Press.
Joyce, J. 1998. *Foundations of causal decision theory*. Cambridge: Cambridge University Press.
Kavka, G. 1983. The toxin puzzle. *Analysis* 43(1): 33–36.
Koons, R. 1993. Faith, probability and infinite passion: Ramseyan decision theory and Kierkegaard's account of Christian faith. *Faith and Philosophy* 10: 145–160.
Kornhuber, H.H., and L. Deecke. 1965. Hirnpotentialänderungen bei Willkürbewegungen und passiven Bewegungen des Menschen: Bereitschaftspotential und reafferente Potentiale. *Pflügers Archiv* 284: 1–17.
Kornhuber, H.H., and L. Deeke. 2012. *The will and its brain*. Lanham: University Press of America.
Kreps, D. 1979. A representation theorem for 'Preference for Flexibility'. *Econometrica* 47(3): 565–577.
Liu, F. 2011. *Reasoning about preference dynamics*. Dordrecht: Springer.
Lukes, S. 1997. Comparing the incomparable: Trade-offs and sacrifices. In *Incommensurability, incomparability and practical reason*, ed. R. Chang, 184–195. Cambridge, MA: Harvard University Press.
Millgram, E. 1997. *Practical induction*. Cambridge, MA: Harvard University Press.
Oakesford, M., and N. Chater (eds.). 1999. *Rational models of cognition*. Oxford: Oxford University Press.
Ramsey, F.P. 1931. Truth and probability. In *The foundations of mathematics and other logical essays*, ed. R.B. Braithwaite, 155–198. New York: Harcourt, Brace & Co.
Raz, J. 1999a. Agency, reason and the good. In *Engaging reason*, ed. J. Raz, 22–45. Oxford: Oxford University Press.
Raz, J. 1999b. Explaining normativity: On rationality and the justification of reason. In *Engaging reason*, ed. J. Raz, 67–89. Oxford: Oxford University Press.
Reisner, A. 2011. Is there reason to be theoretically rational? In *Reasons for belief*, ed. A. Reisner and A. Sieglich-Petersen, 34–54. Cambridge: Cambridge University Press.
Richardson, H. 1986. *Practical reasoning about final ends*. Cambridge: Cambridge University Press.
Rippon, S. 2011. In defense of the wide-scope instrumental principle. *Journal of Ethics and Social Philosophy* 5: 2.
Sen, A. 1982. *Choice, welfare and measurement*. Cambridge, MA: MIT Press.
Sen, A. 1985. Well-being, agency and freedom. *Journal of Philosophy* 82(4): 169–221.
Sen, A. 1999. *Commodities and capabilities*. Oxford: Oxford University Press.
Sen, A. 2002. *Rationality and freedom*. Cambridge, MA: Harvard University Press.
Skyrms, B. 1990. *The dynamics of rational deliberation*. Cambridge, MA: Harvard University Press.
Sobel, J.H. 1998. Ramsey's foundations extended to desirabilities. *Theory and Decision* 44: 231–278.
Sobel, D., and D. Copp. 2001. Against direction of fit accounts of belief and desire. *Analysis* 61(1): 44–53.
Stocker, M. 1990. *Plural and conflicting values*. Cambridge: Cambridge University Press.
von Neumann, J., and O. Morgenstern. 1944. *Theory of games and economic behavior*. Princeton: Princeton University Press.

Chapter 4
Autonomy and Rational Deliberation About Ends

1 Understanding Means and Ends

The previous chapter contains all the background scenery against which the development of my account of ends-deliberation will play out. With the stage set, I am now in a position to begin that account. The first step is to make our talk of ends more precise, and to characterize them using our theoretical apparatuses of probability distributions and preference-rankings. I take means to be actions, and take an agent's set of potential ends to be the set of outcomes which it is possible for the agent to realize. This complete set of outcomes is ordered in the agent's preference-ranking. For an outcome to count as one of an agent's actual ends, the agent must strictly prefer its occurrence to its non-occurrence. If the agent strictly prefers the non-occurrence of an outcome to its occurrence, the avoidance of that outcome would be one of his ends. Actions can be ends in themselves, insofar as the fact that an action is or has been performed is an outcome of performing that action. If the agent is strictly indifferent between an action's being performed or not, the agent sees that action as a mere means. What we want is to find a way to characterize an agent's means and ends, in terms of the his preference-ranking and probability-distribution, in a way that makes clear whether the agent considers the performance of an action to be an end in itself, or merely considers the action a means to an end. We will then be able to see the agent's preferences and probability judgments as encoding a system of ends and means, and understand deliberation about the adoption of a preference-ranking as deliberation about ends in the fullest sense.

An action which is a means to an end is a causally efficacious way of bringing about the realization of that end. If an action a is a means to end e, then $p(a \mathbin{\square\!\!\rightarrow} e) > p(e)$. That is to say, the probability that e would occur were a to be performed is higher than the simple probability that e will occur. If action a is a merely instrumental means, we can fully characterize it in terms of its relation to its end in both the agent's preference-ranking and his probability distribution. This will be the case if $e \succ -e$ *and* $a \sim -a$ *and* $p(a \mathbin{\square\!\!\rightarrow} e) > p(e)$. If a and b are both merely instrumental means to some end e, we can say that $p(a \mathbin{\square\!\!\rightarrow} e) > p(b \mathbin{\square\!\!\rightarrow} e)$ to express the

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_4

fact that a is a more effective means to e than b is. If a is the best instrumental means to e, we will say: $p(a\Box\rightarrow e) > p(e)$ *and* - $\exists\ x$ *s.t.* $p(x\Box\rightarrow e) > p(a\Box\rightarrow e)$. (Recall that the best means *simpliciter* is the one with the highest expected value, calculated by taking the product of the value of the end and the probability of achieving the end given that the means is taken. This allows us to compare means to distinct ends). We can express the basic idea that an action a is itself a final end, and not a means, with: $a \succ -a$ *and* $\forall e \neq a : p(a\Box\rightarrow e) \leq p(e)$. Here, a-ing is preferred to not a-ing despite the fact that this action does not increase the probability of attaining any other end. We view some of our actions, however, as both ends in themselves and as means to other ends. We can then express the idea that a is valued both as an end in itself and as a means to a further end with: $\exists e \neq a$ *s.t.* $e \succ -e$, $a \succ -a$ *and* $p(a\Box\rightarrow e) > p(e)$. Here, a-ing makes some other end more likely, and attaining that end in addition to a-ing is preferred to attaining that end without a-ing, which implies that a-ing is itself preferred to not a-ing and thus that performing a is valued in its own right. On the other hand, *not* performing some action may itself be one of the agent's ends, even if that action is a means to another of the agent's ends: $-a \succ a$, $e \succ -e$, *and* $p(a\Box\rightarrow e) > p(e)$.

Merely instrumental means are always taken for the sake of achieving their end. But an action which is both an end in itself and a means to some other end need not be pursued for the sake of that to which it is a means—the agent may deem it more valuable than whatever it is a means to. We can express the idea that a is not performed for the sake of end e despite being a means to it with: $e \succ -e$, $a \succeq e$, *and* $p(a\Box\rightarrow e) > p(e)$. Here a-ing is valued in its own right, is a means to e, but is also valued at least as highly as e. If action a is valued as an end but is also pursued for the sake of achieving some other end e, we will instead say: $e \succ a \succ -e$, $a \succ -a$, *and* $p(a\Box\rightarrow e) > p(e)$.

Our goal is a framework for deliberation about what preferences to have—what to value, what ends to pursue—in the first place. To get closer to achieving it, we must enrich our framework by adding the notions of *ultimate* ends and *constitutive* means.

2 An Aristotelian Theory of Ends

I take it that the ultimate end of life is to lead a good one. Here I mean "good" to be taken as a very thin notion. I begin with no assumptions about what the content of a good life is. It may be a virtuous life, or an optimally pleasant one, or one lived in strict accordance to the moral law, etc. We can characterize this thin ultimate end of leading a good life (call it G) in the following way: $\forall x\big((x \succ -x) \rightarrow p(G \mid x) > p(G \mid -x)\big).\&.-\exists y \neq G\, s.t. \forall x\big((x \succ -x) \rightarrow p(y \mid x) > p(y \mid -x)\big)$. The ultimate end is simply that to whose realization every other end would contribute, and there is nothing other than the ultimate end to whose realization every other end would contribute. But we must be careful about the sense in which realizing other ends is a means to realizing the ultimate end of leading a good life. "A good

life," as Michael Stocker observes, "is not a single value" (Stocker 1990, p. 181). The other final ends whose realization is a means to the ultimate end of leading a good life are not *instrumental* means to that ultimate end. There is no state distinct from them, a state of leading a good life, which is linked by nomological causation to their realization. Rather, they are *constitutive* means to that ultimate end—to lead a good life just is to lead a life in which one pursues and attains what is of value. Pursuing and realizing enough of those other final ends constitutes leading a good life, and so realizing one of them is a part of (and in that sense a means to) leading a good life. This is why we express the relationship between an end x and G using an evidential conditional probability rather than a causal one.

We begin, then, with an ultimate end in life, but this end is so thin as to be devoid of any particular content. To deliberate about ends, in the sense that we are interested in, is to deliberate about which final ends one should select as the constituents of that ultimate end—to deliberate about which ends one should pursue in one's attempt to lead a good life. The result of such a deliberation is that one or more ends which one could pursue are ranked above one or more other ends in one's preference-ranking. As the results of one's ends-deliberations accumulate, one's ultimate end starts to 'thicken.' The process of choosing which ends to pursue is the process of filling in a conception of leading a good life—and thus, of giving content to the thin ultimate end one with which one began: "When happiness is grasped by an agent, his deliberative task is to find the particular (type of) action that constitutes it in the actual circumstances. Once found, it plays…[the role of] a concrete specification of his end—a filling out of what happiness consists in" (Reeve 2013, p. 234). A fully worked-out conception of a good life can be seen as an aspirational goal of a life guided by ends-deliberation. My account for ends-deliberation, therefore, will not assume a full conception of the good-life at the outset. This is something that is pieced together along the way, and whether it is desirable or even possible to complete such a conception is not a question I need address here.

The next element in the framework is the Aristotelian notion of a constitutive means. A brief look at Aristotle's theory, through the lens of John Cooper's careful exegesis and interpretation, will help to identify the role played by this notion in a theory of ends-deliberation, as well as illuminate the thoroughly Aristotelian character of my project.

One of Aristotle's familiar assertions is that "We do not deliberate about the ends but about what bears on the ends" (*Nicomachean Ethics* [hereafter *NE*] 1112b11–12). It is surprising, then to claim that Aristotle had a theory of ends-deliberation. To see that he does, Cooper draws our attention to two important points. The first is that Aristotle admits that we have many ends, and allows that any end may be pursued for its own sake, though we must also say that it is pursued for the sake of the ultimate end (Cooper 1986, p. 16). Cooper thus asserts that "what is an end in one practical context, and so not deliberated about, is a means in another, where it is a subject of deliberation" (Cooper 1986, p. 15). We deliberate about ends insofar as we deliberate about which of two ends is a better means to a higher end which is attainable through either, and insofar as we choose among conflicting ends by determining which of the conflicting ends serves the highest end (Cooper 1986, p. 18).

Thus far, it may sound as if Aristotle can only accommodate ends-deliberation in the uninteresting sense discussed above, which can be modeled without any significant expansion of the standard Bayesian model. But we can resist this conclusion so long as we understand precisely what Aristotle includes under the heading of "means." The Greek phrase that is normally so translated is "τὰ πρὸσ τὰ τέλη," literally, "the things toward the ends." The phrase "covers more than just 'means'; it signifies 'things that contribute to' or 'promote' or 'have a positive bearing on' an end" (Cooper 1986, p. 19). In particular, included in the Aristotelian class of means are constituent parts of an end—thus the figures which form parts of a larger figure being constructed are "means" to the end of constructing that larger figure (Cooper 1986, p. 20)—and definitions of what exactly an end consists in—thus providing a definition of what health consists in is a "means" toward the doctor's end of making a patient healthy (Cooper 1986, p. 21).

Constitutive means can be given a probabilistic interpretation using evidential decision theory, rather than causal decision theory. Given that an agent has achieved a part of his end, our confidence that he will achieve the whole of the end should be higher than it was before he achieved that part. The characterization of final ends that are also pursued for the sake of some further end given above assumes that the first end is an instrumental means to the second. If it is a constitutive means instead, we can say $\exists a \exists e\, s.t.\, Pae$ *and* $e \succ -e$ *and* $a \succ -a$ *and* $p\left(e \mid a\right) > p\left(e\right)$ (where Pxy is the mereological relation of parthood).[1] Aristotle's third kind of means is the *determinate* means, which involves making a sharper determination of what one's end consists in. In my model, taking a determinate means will correspond to updating one's preferences after engaging in specificational reasoning. We may take the fact that an agent has done this as evidence that he is at least *slightly* more likely to achieve his end, now that he has a sharper notion of what it is.

That Aristotle understands means to ends in this expanded sense brings his theory closer to the sort of theory of ends-deliberation that I will develop. But even if he can allow for deliberation about what end to adopt *qua* constituent part of some already fixed and fairly specified end, or about how to further specify what an already fixed end consists in, it does not necessarily follow that he will admit the possibility of deliberating about which final ends one should adopt as the basic constituents of the ultimate end—about how one should begin to thicken this thin notion. Can his theory allow even for this? According to Cooper, the answer is: yes and no. There are two distinct questions here. First, does the practically wise agent (φρόνιμος) arrive at his conception of a good life on the basis of deliberation; and second, can the *phronimos* give something like a deliberative chain of reasoning that supports his conception, even if his commitment to it is not based on that reasoning (Cooper 1986, p. 64).

[1] When a final end is a constitutive means to a further final end, the parthood relation allows us to distinguish the constituent means from the end it partly constitutes. Here, the fact that part of an end has been attained raises confidence in the whole end's being attained. Constitutive means may also be instrumental means, insofar as achieving part of an end may have a direct impact on one's ability to achieve the other parts.

On Cooper's interpretation, Aristotle answers the first question in the negative—but for an idiosyncratic reason that we need not adopt. Aristotle understands scientific reasoning as proceeding from indemonstrable first principles of which the scientist must have intuitive knowledge (νοῦ ς) (Cooper 1986, p. 65). Aristotle's refusal to allow that the *phronimos*' particular conception of the ultimate end is arrived at through deliberation is derived from his treating deliberation as the practical analogue of scientific reasoning: what the ultimate end consists in must be known intuitively, and then what lower-order ends promote it, and what those lower-order ends consist in, can be demonstrated through deliberation (Cooper 1986, p. 65). In addition to deliberation, however, Aristotle also emphasizes the importance of *inductive* reasoning, reasoning which begins from facts of observation and leads back to the first principles of the sciences (Cooper 1986, p. 67). When the *phronimos* does moral philosophy, he begins with a thin notion of the ultimate end, and then provides reasons which are meant to convince others that the particular conception of a good life that he has grasped is the one that they should adopt (Cooper 1986, p. 68). So Aristotle does allow that the constituent final ends of the ultimate end can be reached through some sort of process of reasoning.

There is an alternative interpretation of Aristotle's theory of practical reasoning, however, which brings it even closer to the theory of ends-deliberation I am developing. Jonathan Barnes has argued that Aristotle does not take *nous* to be a faculty of intuitive grasping or knowing; rather *nous* is the *state* (ἕξις) of *thoroughly understanding* the starting-points (ἀρχαί)—the truths which cannot be deductively demonstrated because they stand at the beginning of every deduction—of a science. We arrive at this state not through an exercise of intuition, but through observation, experience, and ordinary inductive reasoning (ἐπαγωγή) based on these (Barnes (1994) pp. 267–269).[2] The starting point of ethics is a correct conception of the life

[2] More precisely, we should recognize *nous* as a faculty—capacity, power (δύναμις)—though not one of intuitive grasping or knowing. Rather, it is the faculty of thought—the ability to reason with and about concepts that stand in for universals, not just about particular things. Every (normal) human being has the first potentiality of this capacity—the potential to acquire *nous* in the sense of a state of thorough understanding of starting-points (true propositions employing correct universal concepts). We make this acquisition *via* induction. This state, once acquired, is the second potentiality/first actuality of the capacity. Its second actuality is the activity (ἐνέργεια) of *thinking* (νόησις). The one who possesses the state of understanding is thereby *disposed* to think *correctly*, even when he is not actively thinking—just as the one who possesses habituated natural virtues of character is thereby disposed to desire correctly even when he is not actively desiring, and the one who possesses *phronesis* is disposed to deliberate correctly even when he is not deliberating. This is why *nous*, in the sense of the state of understanding, is an intellectual virtue. It is the excellent dispositive state of the thinking part of the soul, just as *phronesis* is the excellent dispositive state of the calculating part, and the character virtues are the excellent dispositive states of the appetitive part. Knowledge of starting-points is actually the result of a two-step process: the *archai* are framed on the basis of induction, and then revised, refined, and thoroughly grasped through the use of dialectic to solve puzzles (ἀπορίαι), in the form of conflicts with credible common beliefs, which seem to arise from their formulation. The puzzles are solved when the formulations of the *archai* are brought into coherence with the most credible of the common beliefs. See (Reeve 2013, pp. 139–141). This dialectical stage has an analogue in my theory of practical reasoning as well: the revision of one's deliberative choice of ends on the basis of the results of one's ethical

well lived (εὐδαιμονία). So on Barnes' interpretation, Aristotle does take the *phronimos* to arrive at genuine understanding (*nous*) of the basic constituents of *eudaimonia* through a process of practical inductive reasoning—precisely the sort of process I will endeavor to model formally.

If this interpretation is right, how can we square it with what Aristotle says about deliberation being concerned not with the ends themselves, but with the things toward them? The key is to observe that for Aristotle, the inductive reasoning that leads to understanding about practical matters is *not* a form of deliberation. The reason for this is that deliberation (βούλευσις) ends in choice (προαίρεσις)—where there is no choice to be made, there can be no deliberation. And according to Aristotle, the *phronimos* does not choose *a* correct conception of the good life; rather, he *discovers the one* correct conception of the good life. Aristotle's eudaimonism is monistic. The one completely eudaimonistic life for a human being, as we learn in the tenth book of the *Nicomachean Ethics,* is a life of (1) preparing for a career as a statesman; then (2) serving the *polis* as a statesman; followed by (3) a monastic retirement, spent studying, contemplating and discussing astronomy, mathematics, and especially theology as a largely self-sufficient member of a small community dedicated to this activity. This is the life of developing and exercising practical and theoretical wisdom. It is the complete life of complete virtue.

The difficult point in correctly interpreting Aristotle's view is recognizing that, according to him, understanding the starting point of ethics—having the correct conception of *eudaimonia*—is not a function of *phronesis*, the virtue expressed through excellent deliberation. Rather, it is a function of *nous*, which (along with knowledge, ἐπιστήμη) is part of theoretical wisdom (σωφία). Ethics begins not with practical wisdom in the sense of *phronesis*, but rather with *theoretical understanding of practical matters*. As C.D.C. Reeve observes: "[T]he practical sciences have their theoretical side: [*Posterior Analytics*] I.34 89b9 mentions *ta êthikês theôrias* ('theoretical ethics') in the same breath as *ta phusikês theôrias* ('theoretical natural science'), while [*Eudemian Ethics*] I.6 1216b36–39 reminds us that political scientists should have 'the sort of theoretical knowledge (*theôrian*) that makes evident not only the fact, but also the reason why" (Reeve 2013, p. 104). Reeve explains that while Aristotle defines theoretical wisdom in terms of its concern with those things that cannot be otherwise, he does not limit it to those *strictly* theoretical sciences which study necessary existents—astronomy, mathematics, and theology. He applies it to the sublunary natural sciences, and to ethical and political theory as well (Reeve 2013, p. 103). *Nous* is one state of the rational part of the soul with two functions—one strictly theoretical, the other more practical (*De Anima* III.10

deliberations about how to act in particular circumstances (a topic we will come to on Chap. 14). To introduce one further wrinkle: Terrence Irwin has argued (persuasively, to my mind) that Aristotle's conception of dialectic in the *Analytics* is insufficient for the metaphysical realist interpretation he gives to his conclusions, and thus that he does require a doctrine of intuition in the early works. However, he can do away with intuition once he revises his view of dialectic in the *Metaphysics*—although his metaphysical realism is still undermined by an unjustifiably naïve empiricism (Irwin 1987, ch. 7–9). Since I am not a metaphysical realist, this distinction between forms of dialectic does not concern me.

433a13–30 cited in Reeve 2013, p. 232). This view is justified because there is a sense in which theoretical ethics and politics are concerned with what cannot be otherwise—although conditionally, rather than unconditionally so. Man would not be man if his *telos* were other than what it is; he would be a different kind of creature altogether.

It is therefore Aristotle's eudaimonistic monism—his understanding of the process of forming the correct conception of the good life as the making of a discovery—which leads him to assert that *phronesis* and deliberation are not concerned with the end itself. These assertions must not be taken to indicate that Aristotle denies that we arrive at that correct conception through a process of practical reasoning. His monism is the basis for distinguishing that sort of inductive practical reasoning from deliberation. It is also the basis for his view of deliberation as exclusively concerned with actions, to the exclusion of non-action outcomes. The outcome which is the end of man is: having lived the kind of complete life of complete virtue described above. Deliberation is then exclusively concerned "to find the particular (type of) action that constitutes [*eudaimonia*] in the actual circumstances" (Reeve 2013, p. 234).

The main point of departure from Aristotle for my theory of ends-deliberation is that it is explicitly and vigorously *pluralistic*. Because I do not share Aristotle's severely limiting metaphysical-biological assumptions about the proper end of a human life and the proper function of a human being, I recognize that any individual who has reasoned as well as he could about how to fill in his conception of his ultimate end may find that he has a number of equally well-supported options to choose from (or may have found that the path to the conception he develops was one of several equally well-supported paths he could have taken). This turns reasoning about how to fill in one's thin conception of one's ultimate end into a form of deliberation—it makes it a form of reasoning that is concerned with making a choice. And this choice now concerns not only actions which are ends in themselves, but ends which are non-action outcomes as well. We will see in Chaps. 13 and 14, however, that even the agent who has deliberated excellently about what ends to pursue may find himself in a particular situation in which taking the most effective means to achieving that end, or even any means to it at all, is not what he ought to do, all things considered. This latter judgment about what ought to be done in a particular case is arrived at through *ethical deliberation*, with which ends-deliberation will have to be reconciled. It is this ethical deliberation which is the closest thing in my overall theory of practical reasoning to Aristotle's conception of deliberation—though I will model it as an inductive process like ends-deliberation, rather than as deductive or syllogistic. My theory will end up integrating ethical deliberation not only with ends-deliberation, but with means-deliberation, mere instrumental reasoning, as well. This is likewise necessitated by my pluralism as opposed to Aristotle's monism. On my view, ethical deliberation turns out not to be concerned with what to do, so much as what not to do—with setting the constraints within which we may pursue our goals. Aristotle's view, by contrast, is that good ethical deliberation determines the specific action in a specific set of circumstances which

is a constituent of *eudaimonia* correctly conceived—there is no space for further action-oriented reasoning.[3]

The one wrinkle in this interpretation of Aristotle is the fact that there are passages in which he seems to indicate that it is possession of the virtues of character, not any sort of practical reasoning, which is responsible for the fact that one has the correct conception of the end: "Virtue sets the target, practical wisdom the things toward it" (*NE* VI.12 44a8–9). Reeve comments:

> If the target someone aims at is noble, his character must be virtuous and thus praiseworthy, since virtue of character is what ensures the correct deliberate choice of noble and praiseworthy ends. (Reeve 2013, pp. 249–250)

One reason to be immediately suspicious of this interpretation of Aristotle's claim is that it contains an error in modal logic. The first half of Reeve's sentence is supposed to be entailed by the second. Reversing the order, we can rewrite Reeve's argument as: (1) If one's character is virtuous, then necessarily, one makes the correct deliberate choice; therefore, (2) if one makes the correct deliberate choice, then necessarily, one's character is virtuous. (1) is, for Aristotle, perfectly true; but (2) does *not* follow from it. So at the very least, we need to look for some other reason for Aristotle to endorse it before we attribute it to him.

The ready answer, and one which seems to square with what Aristotle says about practical inductive reasoning, is that the evidence which is used in that reasoning is the sensory experience of pleasure or pain—or rather, the "practical appearance" (φαντασία), the propositionally structured representation of that sensation (αἴσθημα) of pleasure or pain, which mediates between sensation and *nous* and is fit to serve as data for inductive reasoning—that accompanies performing a given action for a given agent.[4] The agent who possesses the virtues of character—that is to say, natural virtue reinforced by habit, but not yet practical wisdom—is already experiencing pleasure and pain with respect to the appropriate actions. These experiences provide just the data he needs to arrive inductively at the conclusion that the end promoted by the actions he finds pleasant is the correct end. Once he has acquired practical *nous*, and has a precise understanding of an explicitly formulated conception of the end, he is in a position to determine *via* deliberation what action is the virtuous one in any specific situation. This is what it is to be practically wise. The one who does not have the same experiences of pleasure and pain lacks access to the same reliable data as the naturally virtuous. The virtues of character, moreover,

[3] Another difference is that for Aristotle, excellent deliberation and mere instrumental reasoning—the latter being reasoning about the most efficient way to accomplish any proposed end, whether good or bad—are the same in form. The latter expresses the state of cleverness (δεξιότης), but (obviously) not practical wisdom, while the former expresses both (Reeve 2013, p. 250). In my theory, ethical deliberation (the closest thing to Aristotle's own view of the *phronimos*' deliberation) has an inductive structure, while instrumental reasoning is represented as in classical decision theory. So all three forms—ethical, instrumental, and ends-deliberation—will have to be integrated into a single coherent multi-stage process of practical reasoning.

[4] There are many views on the core meaning of Aristotle's notion of *phantasia*, but this is the one I endorse. See (Frede 1995, pp. 290–294).

as states of the non-rational part of the soul, are what shape one's appetites (ἐπιθυμίαι), the impulses that belong to the non-rational part of the soul. Action is the result of the combination of appetite and wish (βούλησις), the latter being the corresponding impulse of the rational part (Reeve 2013, p. 206).[5] Virtue of character is thus required for one's appetite to be directed at the right end. And, since wish follows practical understanding (*nous*) in being for whatever end the agent judges to be good, virtue of character also plays an important role in developing correct wish, by giving the agent access to reliable data for the process of inductive reasoning that leads to genuine practical understanding (*NE* V.9 1136b7–8). Virtue of character, therefore, normally and most straightforwardly enables one to arrive at the correct conception of *eudaimonia*, and moves one to perform whatever action one has determined, through deliberation, constitutes that end in a particular case. This is Aristotle's view.

None of this, however, amounts to the conclusion that virtue of character is *required* for one to have a correct understanding of the end, to make correct deliberate choices, or to act on those choices. And indeed, we should not reach this conclusion; this *cannot* be Aristotle's view. The reason is that it is flatly inconsistent with what Aristotle says about the phenomena of strength and weakness of will. Here is Reeve, some 30 pages earlier, in a very clear passage which, unfortunately, contradicts the one just quoted:

> Since an incontinent [i.e. weak-willed] person's wish is for the correct target, namely, happiness correctly conceived, the 'best thing, the starting point, is preserved in him' (*NE* VII.8 1151a24–26). But…his appetites…are not in a mean…[H]is 'deliberate choice is good' (*NE* VII.10 1152a17) since it is for the correct end…incontinent people (unlike children or beasts) can deliberately choose. (Reeve 2013, pp. 218, 219, 221)

The strong-willed agent, moreover, manages even to perform the right action, since his good wish overcomes his bad appetites. It must be possible, therefore, for one who does not possess a virtuous character—one who is not already moved to pursue the correct end, who does not already take pleasure and pain in the right things—to engage in the process of practical inductive reasoning that leads to a genuine practical understanding of the correct end. More difficult, less straightforward—but possible nonetheless. There must be other forms of suitable evidence to which he does have access. We will see that this conclusion fits nicely with my theory of ends-deliberation, in which one's emotional responses to pursuing and achieving potential ends are but one type of evidence among several. We are fully justified, then, in viewing my theory of ends-deliberation as an attempt to formalize a pluralistic version of Aristotle's own account of practical reasoning.

[5] See also (*De Anima* III.9 432b26–433a8).

3 A Formal, Endogenous, Dynamic Model of Rational Deliberation about Ends

3.1 Preferences, Evidence, and Updating

One of the basic Aristotelian commitments of my project is that autonomy is a capacity which an agent exercises by deliberating about what ends to pursue:

> *Agent Autonomy*: An agent is autonomous insofar as he exercises the capacity of autonomy, the core component of which is the capacity for deliberating about ends. Ideal competence in exercising this capacity is the hallmark of the Aristotelian *phronimos*.

As we set out to represent ends-deliberation, the first point to keep in mind is that in deliberating about his ends, an agent is not engaged in a very different sort of reasoning than he is when he is deliberating about instrumental means. The ends-deliberator has some courses of action available to him. He has the opportunity to adopt one preference-ranking rather than another. He has as many such courses of action open to him as there are ways for him to rank potential ends. He is trying to decide which choice is best, which ends are more valuable for him than which other ends, given his particular situation.

In choosing a preference-ranking, the agent chooses a system of ends to pursue. But we have seen that we can just as well describe him as choosing a system of means. The systems of ends he has to choose from are systems of constitutive means to his ultimate end. So it should not be surprising that a formal model for representing reasoning about instrumental means, like Ramsey's, can be adapted to represent reasoning about ends. The agent's conception of the ultimate end, however, is a thin one: it is just the end of leading a good life. So the ends-deliberator cannot choose which constitutive means is best based on how great a contribution that means will make to his ultimate end. He does not know what his ultimate end is under any description that would allow him to make that sort of determination. The process of filling in the content of his conception of the ultimate end is itself the process of selecting final ends which are constitutive means to that ultimate end. So as Millgram has suggested, the ends-deliberator should proceed by gathering evidence that one end is more valuable than another, no matter his specific ends may turn out to be.

Suppose an agent is trying to determine which of two potential ends a and b is better for him to pursue. There are three preference judgments he could make: $a \succ b$, $a \sim b$, $b \succ a$. To begin deliberating, the agent needs to assign both a subjective probability to the content of each of these judgments: to assign a subjective probability of p_1 to a being superior to b; a probability of p_2 to a and b being equally good; and a probability of p_3 to a being inferior to b. What will enable him to do this? As he begins to fill in his conception of a good life, the ends-deliberator must start from somewhere. The probabilities he assigns to these initial preference judgments must be derived from some source. This requires an assumption about the deliberating agent. Agents do not begin to deliberate about ends from a state of impartial detachment. They begin with pre-deliberative attachments to at least some of the ends

available to them. These initial attachments provide the raw inputs for the process of ends-deliberation.

It may very well be that the biggest obstacle to a precise account of ends-deliberation heretofore has been a failure to appreciate the importance of this point. Ends-deliberation is not the process of formulating a set of reflectively endorsed ends from out of thin air. It is, rather, the process of moving from a set of ends which one finds oneself with, but does not reflectively endorse, to a new set of ends which one does reflectively endorse. The whole challenge lies in representing a process which could justifiably be thought to result in a set of ends that would merit reflective endorsement. It most decidedly does not lie in finding a way to generate rational and deliberate fundamental preferences *ex nihilo*. That is, in all likelihood, impossible; and thus it may be that pessimism about the very idea of ends-deliberation derives from the misconception that such deliberation would have to amount to that particular impossibility. Whatever the right story is about how human agents acquire their initial attachments, it is sure to be a complex biological, psychological, and socio-cultural one. It is here that something like the feature-based model of preference of Dietrich and List discussed in the last chapter, with its focus on attention shifts and motivational salience, likely has a role to play. But it will not be a story of deliberation about one's preferences. That sort of deliberation only enters the scene when the agent begins to question whether he is justified in having the preferences he has.

These initial pre-deliberative attachments take the form of both first- and second-order preferences. Thus far, I have characterized first-order preferences as judgments about what is welcome, or worthy of choice. But initial, pre-deliberative first-order preferences are not judgments; they are representations of affective attachments. They are the sort of things which an agent may simply find himself with. The first step in the process of deliberating about one's ends is to come to see one's initial preferences as judgments about what is choiceworthy. With this development of perspective comes the realization that one hopes that the preferences one has are the ones one should have. This hope is a second-order preference. The agent who prefers a to b, and is considering different ways that his world could turn out to be, will hope that he learns that a is more choiceworthy than b, and be glad if he does learn that this is so. This second-order preference reflects the agent's own affective attachments—the first-order preference-ranking which initially sits at the top of this second-order ranking just is the agent's pre-deliberative first-order ranking—and tracks his degree of attachment to his first-order preferences themselves, in the same way that the first-order preference tracks his attachment to his ends. He is attached to a, and a discovery that a really is the superior option would be followed by a greater feeling of satisfaction than would follow the other possible discoveries.

A full set of second-order preferences will constitute a meta-preference-ranking R over possible preference-rankings R_1, R_2… This is precisely what Sen has suggested as the starting point for an account of autonomous deliberation. Let us spell this out in some detail. Suppose there is a set $E = \{e_1, e_2, \ldots\}$ of ends-propositions. Each e_i is a proposition which denotes a possible end that an agent might seek to

attain. A possible preference-ranking R_i is a proposition which asserts that the ends in E stand in a particular sequence of preference relations. The set $\mathbf{R} = \{R_1, R_2, \ldots\}$ is the set of all such preference-rankings. The meta-preference-ranking R is a proposition which asserts that the preference-rankings in **R** themselves stand in a particular sequence of preference relations. From the perspective of R, the R_i's are themselves prospects, from which a choice must be made. The particular preference ranking R_m which initially tops the meta-ranking R represents the agent's actual initial pre-deliberative attachments—the first-order preferences which the agent (initially) hopes to discover are the ones he ought to have.

Since R is neither more nor less than a preference-ranking over proposition-prospects, we can assume that it obeys the axioms for a Ramsey-style evidential decision-theory as stated by Bradley. We can then represent R with a desirability function, where $des(R_i)$ denotes how welcome the news that R_i is the preference-ranking which the agent ought to adopt would be to that agent. We can then also define a probability function, where $p(R_i)$ is the agent's degree of credence that R_i is the preference-ranking which the agent ought to adopt.

We will see now that the process of deliberating about ends should be modeled as the process of choosing what first-order preference-ranking to intentionally adopt on the basis of the subjective probabilities and desirability-values assigned to the rankings in the meta-preference. Most of the work of deliberating about ends—like most of the work of deliberating about means—involves updating the relevant probability judgments in the light of evidence. A rational agent aims to determine what preferences he *ought* to adopt—what preferences are best supported by all the relevant considerations. He will endeavor to gather evidence that bears on the question of which of his possible first-order preferences he should actually reflectively endorse—evidence that one end actually is more (or as) choiceworthy than (as) the other. At the beginning of his inquiry into the choiceworthiness of possible ends, the evidence the agent gathers will only support the thin conclusion that pursuing one end is likely to be more conducive to leading a good life than is pursuing another end. As the agent fills in a conception of a good life by selecting final ends, he will be able to search for evidence that a potential end will make a significant contribution to *his* conception of a good life. Changes to his preferences will become more subtle, and adopting them will involve less drastic corrections to the courses of action he has embarked on, than they did earlier on. Evidence of the value of ends, as best I can tell, comes in seven basic forms: the actualization of physical and mental capacities, emotional responses, achievement, effects on the interests of others, satisfaction of categorical reasons, testimony, and coherence with other ends. Let us discuss these in turn.

The first type of evidence is the degree to which pursuing an end requires that the agent exercise, and thus have developed, his various capacities—physical, psychological and intellectual. The status of such facts as evidence reflects the basic fact that humans lead lives of both the body and the mind, and that in both of these spheres we possess the potential for generation and development. Worthwhile ends are often challenging. And to say that they are challenging is to say that pursuing and achieving them requires that one deploy skills and abilities—such as stamina,

focus, problem-solving, abstraction, inference, etc.—to a relatively high degree. When pursuing one end requires a greater degree of skill and ability than pursuing another, or requires that the agent exhibit several such skills and abilities in concert, as opposed to just one or two, this is good evidence that the one end is superior to the other. For pursuing and achieving such an end involves a greater realization of the agent's potential for self-development. The ability to pursue effectively and achieve such ends is the result of a commitment to personal excellence. We have good reason to think, then, that such ends are part of what constitutes a worthwhile life. Ends which require the exercise of high-level capacities, moreover, are often ends whose active pursuit will elicit positive emotional responses in the agent.[6] This is our next type of evidence.

The second type of evidence is one's emotional responses in the course of pursuing one's ends and contemplating one's potential ends. This reflects the emotional dimension of human life—an aspect of our psychological existence no less important than the intellectual dimension. If the end one is pursuing is choiceworthy, one would expect that the pursuit would be accompanied by positive emotions.[7] There would be feelings of pleasure, fulfillment and pride that accompanied one's efforts to achieve the end. If one were pursuing an end that was not conducive to leading a good life, one would instead expect to be left feeling dispirited and unfulfilled by that pursuit. It is important to note that I am not assuming that beneficial consequences to others or one's own feelings of satisfaction are themselves the ultimate end for the sake of which other ends are sought. Rather, these positive consequences are by-products of the pursuit of an end that is conducive to leading a good life, and are types of evidence that the end pursued is a constituent of a good life. They may, in addition, be constituents of a good life in themselves. It may be that pleasure, self-satisfaction, and furthering the pursuits of others are part of any good life. We need only be careful not to assert that they are the whole of the ultimate end, or that when they result from the pursuit of any other end, they are the point of that pursuit. An individual agent might reach this conclusion in the process of filling in a hedonistic or utilitarian conception of the ultimate end. But arriving at such a conception is no guarantee of my framework, which is neutral with respect to fleshed-out normative theories of the good life.

The third type of evidence that one's ends are worth pursuing is that one manages to achieve at least some of them, and none of them are obviously so far beyond one's potential as to render their pursuit hopeless. We are beings who plan, intend, and pursue. No robust understanding of human life is possible without recognizing this dimension of it. And for any being that plans and pursues, one element of a life successfully led must be the achievement of things pursued. One consequence of pursuing a choiceworthy end, then, should be a fairly full realization of that end. If one is more likely to achieve one end than another given one's circumstances and natural capacities, or to realize one more fully than another, that is evidence that one should

[6] This is the idea behind Rawls' "Aristotelian Principle" (Rawls 1971, p. 326).

[7] For a development of the suggestion that emotional responses are (often) responses to the presence of objective value, see (Nozick 1989, pp. 87–98).

choose to pursue that end rather than the other. And if one's pursuit of an end is going poorly, and has been for some time, despite one's best efforts, this is evidence that the end one is pursuing is less choiceworthy than its alternatives. One leads a good life not just by pursuing choiceworthy ends, but by achieving at least some of the ends one pursues.

The fourth type of evidence is the effects that one's pursuit of one's ends has on others. The status of these facts as evidence for the value of ends reflects the social dimension of human life. If an end one is pursuing really is a constituent of leading a good life, one would expect that others would benefit in some way as a result of one's pursuit of an end, at least if that pursuit is going well. The effects of pursuing a choiceworthy end should, on the whole, aid others in pursuing their ends, rather than hinder and harm their efforts. If one finds that one's pursuit of an end has, on the whole, beneficial effects on others' pursuits of their ends, then it would be reasonable for one to interpret that as some evidence that the end one is pursuing is genuinely choiceworthy, that it really is conducive to leading a good life. If, on the other hand, one's pursuit of an end consistently frustrates others' pursuits of their ends, this should be taken as some evidence that the end one has chosen is not part of leading a good life.[8]

Facts which the agent takes to be categorical reasons for pursuing ends of certain types are the fifth sort of evidence that bear on the question of how valuable an end is for an agent.[9] Whatever view of the basis of normativity one may subscribe to, to lead a human life is to be immersed in normativity—to operate in what Wilfred Sellars calls the logical space of reasons. We pursue our ends through individual actions performed in individual circumstances. When an available action would advance an agent toward his end, the fact that it would do so will be taken by the

[8] Two caveats are needed. First, if one is pursuing an end that can only be realized by one person or group, and is competing with others who are pursuing the same end, the fact that one's pursuit frustrates theirs should not be taken as evidence against the choiceworthiness of the end. Second, in a perverse community in which the majority aims to exploit and oppress the minority, pursuit of a genuinely choiceworthy end will hinder most others' pursuit of their ends. But we need only remember that the presence of the sort of consequences I am discussing is only one type of evidence that an end is choiceworthy, and that this sort of evidence may be contradicted and outweighed by other sorts.

[9] If we really want to be precise, we should say that *apparent categorical reasons* are a type of evidence that bears on this question. E is an apparent reason to ϕ just in case E would, if true, be a reason to ϕ. This is because what the agent treats as evidence is whatever he takes to be a reason for choosing one of his options, and he may of course be wrong in thinking that some or all of the reasons he takes to be present really are so. The agent's judgment $p(E)$ then expresses how confident the agent is that E is true—that a given apparent reason really is present. If ϕ is the proposition that the agent should perform some action, then the value of $p(\phi|E) - p(\phi)$ can be understood as the degree of support the agent takes E to give to ϕ (or, if this is negative, the extent to which it counts against), and $p[p(\phi|E) - p(\phi) = n]$ can be understood as the agent's confidence that he has correctly gauged this degree of support. The agent may, of course, be wrong about both of these points as well, and change his mind about them later on. All of the types of evidence discussed in this section are apparent reasons which bear on the agent's judgments about his possible preferences. Apparent categorical reasons are simply one such type, as categorical reasons are one type of reason. I will have much more to say on the nature of reasons in Chap. 13.

agent as a reason to do it. But there will normally be many other reasons, favoring various actions, which must be taken into account in determining what he has most reason to do. These reasons may derive from his other ends, or they may be categorical. When an agent is first attempting to determine which potential ends will contribute to his thin ultimate end, and thus has no other ends of his own from which to derive reasons, he will still have categorical reasons to consider. These are the reasons that apply to him regardless of what specific ends he has or will eventually adopt. If the reasons that favor pursuing one's end are consistently outweighed by reasons favoring some other action, this should be taken as evidence against the value of pursuing that end. If, for example, pursuing one's end consistently requires lying or breaking promises (actions which we often have categorical reason not to do), and the reasons that favor the action that advances one's end (including the fact that it advances one's end) do not outweigh the opposing reasons in favor of telling the truth or keeping the promise, this is good evidence against the value of pursuing that end. On the other hand, if the balance of reasons consistently supports the action that will advance one's end, this is evidence that that end is conducive to leading a good life.

The sixth type of evidence is testimony. This too reflects the social dimension of human life. One agent can often observe that a second agent's pursuit of his ends involves the exercise of his capacities, results in beneficial consequences to others, consistently leads to the achievement of the end pursued, is usually supported by the categorical reasons that apply in the situations in which he acts, and induces positive emotional responses in him. This is all good evidence for the first agent to think that the second agent is pursuing genuinely choiceworthy ends. This, in turn, is a good reason for the first agent to take what the second agent has to say about what is valuable in life seriously. The testimony of such agents is thus another important source of evidence. Because preference-rankings of ends are position-dependent—what is a constituent of a good life for one agent may not be so for another, due to differences in circumstances and natural capacities—incorporating such testimony into one's judgments is somewhat tricky. I return to this point below.

The final type of evidence that one end is superior to another is how well those ends cohere with the others ends of the agent. The inclusion of coherence on our list of ends may be seen as a nod to Alistair MacIntyre's idea that a good human life must be a life that possesses a certain "narrative unity" (MacIntyre 1985, ch. 15). The fact that we are beings who plan and pursue necessitates including achievement on our list of evidence for the value of ends. But if our plans, pursuits and achievements lack coherence, lack narrative unity, they fail to be intelligible. Taking coherence among ends in a set as evidence for the value of the ends in that set expresses the idea that a good human life should be one which is intelligible both to the one leading that life and to others in his community. But what precisely is coherence among ends? It is something stronger than the consistency that Bratman discusses. When an agent's ends cohere, it is not just the case that it is possible for him to achieve them all. Coherent ends support one another in some way. This sort of coherence among ends is what Richardson appeals to when he discusses resolving conflicts between ends (Richardson 1986, pp. 152–153). Richardson says that ends

cohere when they provide each other with *explanatory* support; that is, when the fact that an agent is pursuing one end helps to explain why he is pursuing another (Richardson 1986, p. 150). But he has little to say beyond that. The main problem is that no one has yet developed a precise way to measure, or even to characterize, coherence among ends. I am inclined to think that a probabilistic measure of coherence among ends could be developed. Given a probabilistic characterization of means, coherence among ends could be taken to be a measure of how probable the agent's attainment of final ends makes his attainment of further final ends to which the ends attained are also means, compared to how probable the attainment of those further ends would have been had different antecedent ends been pursued and attained, or compared with how probable the attainment of different further ends would be. And the higher-ranking are the others ends with which a given end coheres, the stronger is the evidence that that end should be adopted, and should be preferred to other ends which do not cohere strongly with other high-ranking ends.

Recall that our ends-deliberator began with a meta-preference-ranking over possible first-order preference-rankings, and that each element of the meta-ranking was assigned both a desirability-value and a subjective probability. The result of his exposure to evidence that bears on which ends are more valuable than which is that he updates these initial probabilities. This process of probability updating in the light of evidence acquired through experience is the engine of ends-deliberation. Thus do we transform Millgram's basic insight that we learn what matters from experience into a framework for precisely representing the process of deliberating about ends. When the probability of a hypothesis is updated in the light of new evidence, the updating proceeds by means of generalized conditionalization. The main benefit of using generalized, rather than simple, conditionalization is that we need not assume the agent is certain about his evidence. Our ends-deliberator, for instance, may be fairly confident that the balance of reasons in his present situation supports his pursuit of his end, but not certain that it does. When a hypothesis is updated on non-testimonial evidence, updating by generalized conditionalization proceeds as follows (Jeffrey 2004, p. 54):

$$p_{new}(H) = \sum_{i=1}^{n} p(H \mid E_i)\, p_{new}(E_i)$$

Here, the hypothesis concerns the truth, or correctness, of a given preference ranking R_i. The new, updated probability which the agent assigns to the truth of that proposition depends on the old probability and the agent's confidence in his observations of relevant evidence.

Updating by incorporating testimonial evidence is slightly more complicated than updating on other forms of evidence. Suppose *A* is deliberating about whether to pursue a given end. He believes with a certain degree of confidence (p_{old}) that performing an action necessary to achieving that end is supported by the balance of reasons. His friend *B*, who did pursue and then abandon a similar end, initially agrees. After learning more from *A* about *A*'s situation, however, he revises his judgment. He tells *A* that although he (*B*) was also confident at the time that his actions

were supported by the applicable reasons (p'_{old}), subsequent experiences made him more sensitive to his obligations, and led him to believe that pursuing that end conflicted with his obligations. He took that as good evidence that the end was less valuable than he thought (p'_{new}). After learning more about *A*'s circumstances, *B* judges that *A* is under the same sorts of obligations as he was, and so judges that the end is not as valuable for *A* as *A* takes it to be. *A* currently has a high degree of confidence that the balance of reasons supports pursuing his end, but he trusts *B*, and wants to update his probability (to some p_{new}) on the basis of *B*'s advice. He does not, however, want simply to adopt *B*'s revised judgment about how valuable the end is for someone in his circumstances. *A* is more confident that his pursuit is supported by the reasons that apply to him than *B* was in the past that his (*B*'s) pursuit was supported by the reasons that applied to him ($p_{old} \neq p'_{old}$). What he wants to do, then, is to update his current probability on the basis of the change in *B*'s probability effected by *B*'s subsequent experiences, but to ignore *B*'s prior probability (which he takes to be lower than his). He can do this as follows (Jeffrey 2004, p. 56):

$$p_{new}(H) = \sum_{i=1}^{n} \pi'(E_i)\, p_{old}(H \,\&\, E_i)$$

Here, $\pi'(E_i)$ is *B*'s probability factor for E_i, equal to the ratio of *B*'s old and new probabilities on E_i: $p'_{new}(E_i) / p'_{old}(E_i)$.

By updating the initial probability distribution over the rankings in his meta-preference on the basis of evidence, the agent begins the process of constructing a deliberative preference-ranking over actions—the actions of intentionally adopting one preference-ranking or another. Prior to deliberation, the agent has not intentionally adopted any set of ends, and has no preferences over acts of adopting such preferences. These preferences, and the intentional decision to adopt one particular preference-ranking, one particular set of hierarchically structured ends, are the *products* of the process of ends-deliberation. A hypothesis regarding what preference-ranking is correct is updated on the kinds of evidence relevant to that hypothesis. $p(R_1)$ is the agent's degree of credence that preference-ranking R_1 is the ranking the agent ought to adopt. But a judgment that adopting R_1 is superior to adopting R_2—a preference for the action of adopting that ranking rather than another—must be based on more than the relative values of $p(R_1)$ and $p(R_2)$. Since ends-deliberation is a practical activity, the agent's degree of attachment to the preferences he is deliberating over—the relative values of $des(R_1)$ and $des(R_2)$—must also come into play. A deliberative preference for intentionally adopting one preference-ranking over ends rather than another is a total comparative evaluation of the preference-rankings among which one is deciding. As such, it must take into account all considerations relevant to that decision. The types of evidence that inform the agent's probability judgment are certainly among these. But they are not exhaustive of them. Also relevant is the agent's varying degrees of attachment to the preferences among which he is choosing. The facts of those attachments are not themselves evidence of the actual correctness of those preferences. That the agent

wants a particular preference ranking to be the one he ought to adopt does not make it any more probable that it is so, and the rational agent knows this. But it certainly *matters* to the agent how attached he is to a given preference-ranking. The adoption of preferences over ends, after all, serves to structure the agent's life and pursuits, and there is nothing suspect in the claim that the attachments he finds himself with should make a difference to the way he conceives of the structure of his own life, and should play a role in the construction of that conception. It is *his* life, and he is the one who must live it—so he must be able to *live with* the life he has chosen. We wish, moreover, to characterize autonomy, conceived of as excellence in deliberating about ends, as a realizable achievement, as a trait which ordinary human beings are capable (perhaps with difficulty) of exemplifying. Suppose the agent has determined that $p(R_1) > p(R_2)$, but he is nonetheless very attached to R_2, and so $des(R_1)$ is much lower than $des(R_2)$. A conception of practical rationality that requires the agent to adopt R_1 solely on the basis of the value of $p(R_1)$ is too far removed from our purposes to be of interest. We shall see, however, that the agent who deliberates about ends excellently will update $des(R_1)$ in this situation in light of his realization that $p(R_1) > p(R_2)$.

So with "$i(R_1)$" representing the action of adopting or deciding on R_1 intentionally, we would then have:

$$i(R_1) \geq i(R_2) \leftrightarrow des_{new}(R_1) \sum_{i=1}^{n} p(R_1 \mid E_i)\, p_{new}(E_i) \geq des_{new}(R_2) \sum_{i=1}^{n} p(R_2 \mid E_i)\, p_{new}(E_i);$$

or more succinctly:

$$i(R_1) \geq i(R_2) \leftrightarrow des_{new}(R_1)\, p_{new}(R_1) \geq des_{new}(R_2)\, p_{new}(R_2).$$

Since an agent can really only adopt one preference ranking, the agent will adopt the preference ranking R_m that satisfies[10]:

$$\max(des_{new}(R_i)\, p_{new}(R_i)).$$

So our model takes the agent's pre-deliberative attachments into account.[11] But why $des_{new}(R_1)$, rather than $des(R_1)$? The agent has acquired evidence relevant to the

[10] Recall that this formulation does not commit us to the view that the deliberative agent adopts those preferences whose satisfaction would maximize his own well-being in some narrow, subjective sense. Rather, it is merely a quantitative representation of the agent's best judgment of what preferences he should, all things considered, have.

[11] And thus we avoid a problem that Aurel Kolnai has dubbed "the fundamental paradoxy of Practice." See (Kolnai 1961). The problem is supposed to be that when an agent deliberates about ends, he must *weight* the very possibilities that he is supposed to be *weighing*. That is, the extent to which he is attached to the various ends among which he must choose determines how choiceworthy he will end up judging those ends to be. But by separating out the agent's pre-deliberative attachments from his judgments of the choiceworthiness of ends—judgments which are based on evidence drawn from his experience of the world—as we have done here, we can escape this problem. The agent does initially weight his options. But those initial weights do not determine his

question of whether R_1 is in fact a better ranking for him than R_2. In the process of acquiring that evidence, his initial second-order preference for R_2 over R_1 may change. The update rule for the new desirability might be something like this, which is based on what Bradley identifies as the rule for "taste change" (Bradley 2008, p. 229):

$$des_{new}\left(R_n\right) = des\left(R_n\right) + \sum_i \left[des_{new}\left(E_i\right) - des\left(E_i\right)\right] \cdot p\left(E_i \mid R_n\right),$$

where each E_i is a piece of evidence on which the agent also updates $p(R_n)$. Here, we consider both how welcome the truth of each piece of evidence in the partition $E = \{E_1 \ldots E_n\}$ would be to the agent before the agent has his learning experience, and how welcome the truth of each piece of evidence is after that learning experience—the experience in light of which he also forms new probability judgments about the propositions in E. Essentially, we are modeling the effect of changes in the agent's taste for reasons on his evaluative judgments about his potential preferences—the reasons he learns about during the experience that also prompts him to revise his probability judgments about his potential preferences. The move to $des_{new}(E_i)$ results from the same experience as leads to the agent's increased credence in E_i, whose truth bears on his probability judgments about his potential preferences. The difference between $des_{new}(E_i)$ and $des(E_i)$ measures the strength of the change in his taste for the reason E_i—a taste which can be cultivated by seeking out learning experiences which will have this effect. These desirability updates on the reasons in E in turn have an effect on how welcome the agent finds the truth of a preference-ranking. As Bradley demonstrates, the above equation determines the extent of that effect (Bradley 2008, pp. 229–230).

On the other hand, the agent's attitude toward the prospect that a given preference-ranking is the one he ought to adopt might change as a direct result of a change in the agent's credence that it is the one he ought to adopt. In that case, we might have an update rule that looks roughly like this:

$$des_{new}\left(R_n\right) = des\left[R_n \mid p_{new}\left(R_n\right)\right].$$

This update rule violates a condition which Bradley calls "the local independence of preference from belief" (Bradley 2008, p. 233). Here, a change in the agent's belief about a proposition leads directly to a change in how welcome the agent finds the truth of that proposition. This condition is one which Bradley requires in order to prove a number of other theorems about preference change which do not concern us. But as he himself admits, it is hardly a requirement of rationality with any great independent plausibility, and he discusses seemingly commonplace examples in

choice. A heavily weighted option will only win out of it finds support in the agent's judgment, which is responsive to his experiences and observations of the world. And those very experiences and observations may cause the initial weights assigned by the agent to change.

which it seems to be violated (Bradley 2008, p. 233). The ideally rational agent—the truly excellent ends-deliberator—will update his desirabilities on the basis of his probability judgments in precisely this way, and such that his desirabilities are in line with his probability judgments.[12] If the updating of his judgments is not by itself sufficient to trigger an update in desirabilities which brings those desirabilities fully in line with his judgments, he can go about remedying this by seeking additional evidence to strengthen his judgments (which can lead to further updates to his desirabilities), and by purposefully cultivating taste changes with the goal of bringing his desirabilities in line with his judgments. In performing these actions with the goal of bringing his desirabilities in line with his probability judgments, the excellent ends-deliberator takes responsibility for his affective attachments (which these desirability assignments track). These taste changes can be cultivated by seeking experiences which will expose one to reasons that bear on one's preferences.

We must be very careful about how we interpret the biconditional in our rule:

$$i\left(R_1\right) \geq i\left(R_2\right) \leftrightarrow des_{new}\left(R_1\right) p_{new}\left(R_1\right) \geq des_{new}\left(R_2\right) p_{new}\left(R_2\right).$$

Recall that our background decision theory is a Ramsey-style theory. The agent has preferences over propositions expressing different possible preference-rankings, and these preferences suffice for the probability and desirability representation on the right-hand side. But we have not assumed that the agent has preferences over actions, and in particular, we have not stated any set of axioms for preferences over axioms that would secure a representation of the sort of preference given on the left-hand side. This is because we are attempting to model deliberation about ends as a *process*, and the deliberative, reflective decision to adopt one preference ranking rather than others is the final stage in this process. We do not want to assume that the agent has preferences over such actions at the outset. We must therefore supplement the background decision theory with an additional assumption which will serve as a bridge between the representation on the right-hand side and the preference over actions on the left-hand side. But the biconditional does *not* connect the preference over actions on the left-hand side with a *typical* expected value calculation on the right-hand side, as it does in the case of modeling instrumental reasoning. $des(R_n)$ is *not* the desirability of the outcome of the action of adopting R_n—it is the desirability of discovering that R_n is *correct*. And $p(R_n)$ is not the probability that the agent occupies a state from which that action will yield that outcome. It is the agent's degree of credence in the correctness of R_n. $p(R_n)\ des(R_n)$ is the expected

[12] One might think that a problem in created for my model, insofar as the updating of probability judgments on the basis of emotional responses—which the desirabilities in my model track—appears to violate the local independence of belief from preference. But this only creates a problem if a feedback loop is created, in which rising desirability (probability) leads to rising probability (desirability) and so on. We can block this by stipulating that an increase in probability that follows from an event that increases desirability never leads to a further increase in desirability, and an increase in desirability that follows from an increase in probability never leads to a further increase in probability. Such a stipulation is perfectly justified if it reflects empirical psychological facts about human agents, which it seems to do.

value to the agent of the proposition R_n's *being true*—of it being true that R_n is the preference ranking the agent *ought to adopt* .

So the Ramsey-style background decision theory allows a representation of the agent's beliefs and attachments regarding the truth of propositions of the type R_n, but in the context of ends-deliberation, this representation must be connected to preferences over acts of adopting preferences by way of a different bridge principle than is used in the context of instrumental reasoning. This is one of the fundamental differences between deliberating about means and deliberating about ends. The corresponding principle in the context of ends-deliberation is:

(ED): An agent will prefer to intentionally adopt a preference ranking over ends R_1, rather than another such ranking R_2, iff the agent's degree of credence that R_1 is the one he ought to adopt, weighted by his degree of attachment to R_1 being the one he ought to adopt, is greater than it is for R_2. That is to say, he will do so iff the expected value to the agent of R_1's being true is greater than that of R_2's being true.

The argument given above—that both the probability that a given preference ranking is correct, and the agent's attachment to that preference ranking, must play a role in the agent's decision-making process about what preference ranking to adopt—supports employing (ED) as the bridge principle in our model of ends-deliberation. We interpret this principle as a rational requirement, akin to the axioms of the background theory: a rational agent should, insofar as he is rational, prefer the action which this principle recommends. Our reading of the above biconditional is thus that an agent should adopt one preference ranking rather than another *if* he *has* determined that the former is the one recommended by this principle, and *only if* he has done so; and that *insofar as he is acting rationally*, he *will* do so. For brevity's sake, I will refer to the expected value to the agent of it being true that he ought to adopt a given preference ranking as the expected value of adopting that preference ranking, even though this is not in the strictest sense correct.

The model thus represents the process of moving from pre-reflective attachments to the adoption of preferences which the agent reflectively endorses. We can outline the stages in this process thus:

(1) The agent finds himself with a set of pre-deliberative attachments R_1.
(2) In the light of experience, the agent's credence in R_1, $p(R_1)$, changes in accordance with the update rule.
(3) This change in credence leads to a change in the agent's degree of attachment to R_1, $des(R_1)$.
(4) As a result of these changes, it is now the case that for some other possible preference ranking R_2, $des_{new}(R_2)p_{new}(R_2) > des_{new}(R_1)p_{new}(R_1)$.
(5) In accordance with principle (ED), the agent moves from his commitment to R_1 to the adoption of R_2. His new commitment to R_2 is, insofar as it is the result of this process, reflectively endorsed.

Thus the model provides a formal representation of the process of moving from pre-reflective attachments to reflectively endorsed preferences.

Let us return to Richardson's example of the politician running for re-election. The politician's emotional response to his plan to use helping the homeless as a

mere means to re-election is evidence that he should value helping the homeless more highly. Though Richardson does not then describe the politician as incorporating this evidence into a process of genuine ends-deliberation, he does convey the effect of observing this evidence on the politician's non-deliberative attachments. The politician suddenly revises his second-order preference, and becomes someone who would rather that it be the case that helping the homeless but not winning re-election is at least as good as winning re-election but not helping the homeless. The term $des_{new}(R_1)$ stands for the desirability-value that represents the revised second-order preference. In this example, it seems that the first version of the desirability update rule given above is the better fit with the way the politician's reaction is described. If he now engages in ends-deliberation by incorporating that evidence and updating his probability assignment, he may conclude that helping the homeless really is at least as good as winning re-election.

Suppose we alter Richardson's example, so that the politician experiences the same negative emotional reaction, but is so attached to his goal of winning re-election that he does not revise his second-order preference—he remains someone who would rather that it be the case that winning re-election is better than helping the homeless. Such an agent, as I have said above, should not necessarily be labeled irrational for failing to adopt an intention to help the homeless even at the expense of winning re-election. In the context of the example, the politician will need to be swayed by observations of consequences other than his negative emotional reaction, or by categorical reasons, testimony, etc. At some point, however, he will be presented with so much evidence that his preferred preference judgment is wrong that the desirability he attaches to it—the degree to which he wants it to be right—will be overwhelmed. At that point, the only rational thing for him to do will be to form the deliberative preference judgment that helping the homeless is at least as choice-worthy as winning re-election, and to revise his intentions accordingly. In the best-case scenario, the process of acquiring this preponderance of evidence regarding the value of helping the homeless will also affect the politician's second-order preference. He will develop an attachment to the goal of helping the homeless and will not be disappointed by his recognition that doing so is at least as good as winning re-election.

In the previous chapter, I criticized the theory of Dietrich and List for modeling changes in the sets of reasons that underlie agent's preferences exogenously. We are now in a position to see that my model does not suffer from this defect. We can endogenously model the phenomenon of practical reflection, discussed by Dietrich and List, whereby an agent comes to recognize some fact as a normative reason that applies to him and is relevant to his preferences to some extent, or to change his mind about how relevant it is. We can model the change over time of the agent's probability judgments of the form $p(R_m|E_X)$, where E_X is some fact which, if observed, the agent either would or would not take as counting for or against the claim that R_m is the correct preference-ranking, in response to exposure to evidence of a different sort. Suppose the agent currently accepts that $p(R_m|E_X) = n$. We can understand this number n to be the agent's *expectation* of the value of $p(R_m|E_X)$, $exp(p(R_m|E_X)) = \int_0^1 p\left(p\left(R_m \mid E_X\right) = N\right)\left(N\right)dN$ (Jeffrey 2004, pp. 62–67). The

agent can update this expectation on some other piece of evidence E_Y, and adopt a new expectation $exp_{new}(p(R_m|E_X)) = exp(p(R_m|E_X)|E_Y) = n'$, assuming for simplicity that $p_{new}(E_Y) = 1$. The agent has changed his mind, in the light of evidence E_Y, about how probable it is that R_m is correct given E_X—he now has a new judgement of $p(R_m|E_X)$.

So my model has the resources to explicitly represent changes in the agent's views regarding what counts as a reason for a preference and to what extent a given fact would count in favor of a preference, in a way that depends on experiences of other types of evidence. The only phenomena which are modeled exogenously are the occurrence of events and obtaining of facts in the external world, and the agent's experience and observation of those facts and events. This, as I stated at the end of the last chapter, is precisely what we want from our model. The other phenomenon discussed by Dietrich and List of a reason becoming emotionally salient, whereby the recognition of a normative reason for pursuing some end has an impact on the agent's affective attachments, is modeled by the update rules for desirabilities based on the acquisition of new evidence and on updated probability judgments based on that new evidence:

$$des_{new}\left(R_n\right) = des\left(R_n\right) + \sum_i \left[des_{new}\left(E_i\right) - des\left(E_i\right)\right] \cdot p\left(E_i \mid R_n\right);$$

$$des_{new}\left(R_n\right) = des\left[R_n \mid p_{new}\left(R_n\right)\right].$$

Thus, what is pure interpretation of phenomena modeled exogenously in the work of Dietrich and List is represented explicitly and endogenously in my model.

This endogenously modeled process of working out and refining one's views on what sorts of considerations to count as evidence, and how to weight that evidence, can be viewed as a transition from an initial state of agnosticism about ethical theory, to a state of commitment to the truth of some particular ethical theory. Suppose an agent is initially agnostic about ethical theory. He equally countenances the relevance of all those facts and features which consequentialists, Kantians, intuitionists, sentimentalists, and Aristotelians each recognize as relevant to the question of what one ought to do.[13] For example, he allows that there are some facts that are categorical reasons for action, and make certain judgments of the probability that pursuing various potential ends will contribute to his leading a good life in the light of his observations of facts he takes to be categorical reasons; but he simultaneously acknowledges the equal relevance of his observations of consequences, emotional responses, etc., to those same judgments. But now suppose the agent takes a university class on ethical theory, the result of which is that he becomes much more disposed to the work and thought of Bentham. The student takes his understanding

[13] The absence of any mention of "virtue" from the list of types of evidence given above is explained by my understanding of Aristotle's theory of virtuous action as being, essentially, action for the right reason. In Chap. 13, I will argue for a neo-Aristotelian ethical theory which deliberately accepts as normatively relevant all the types of considerations discussed above. We will return to the question of what place virtue should occupy in ethical theory in Chap. 14.

of Benthamite utilitarianism to be relevant to the question of what preferences he ought to have, insofar as these theories both provide answers to the question of what type of reason is relevant to properly determining one's ends, choices, and actions. Suppose for a potential preference-ranking R_m, the agent previously judged that some hedonic consequence would be relevant to the correctness of the ranking in the following way: he held $exp(p(R_m|E_B)) = n$. The experience of taking the course (call it E_C, and assume that $p(E_C) = 1$, i.e. the student is certain what the content of the course was) bears on this expectation (though it is irrelevant to the agent's credence in the obtaining of the relevant fact E_B). It affects the agent's beliefs regarding the relevance of the type of practical reason emphasized by the theory to his own practical choices. In our example, given E_C, the expectation is updated, $exp_{new}(p(R_m|E_B)) = exp(p(R_m|E_B)|E_C) = q$, with $|p(R_m) - n| < |p(R_m) - q|$. The agent has taken a step away from being an agnostic about ethical theory and toward being a hedonic utilitarian with respect to his practical choices as well as his theoretical views.

The decision to represent desirabilities as updated on the basis of updated probability judgments about what one ought to prefer, and to represent agents as choosing ends based in part on these judgments, has implications for an old philosophical debate about motivation. As I mentioned above, I take the desirabilities assigned to possible preference-rankings to track the agent's affective attachments. The degree of an agent's affective attachment to a particular outcome—call it the "affective value" of the outcome—can be taken as the valuation which represents the position of that outcome in the preference-ranking with the highest desirability. These are the agent's first-order preferences prior to deliberation. Following the psychologist and neuroscientist Edmund Rolls, I understand these affective values to be encoded by physical, neurological states—in particular, states of the brain's reward system (Rolls 2005).[14] The Humean Theory of Motivation is a family of views about action, each of which maintains at least the first of the following four theses (or something like them): (1) an agent performs an action only if he has a desire to perform the action, and (2) always acts in a way consistent with a preference for the outcome which he desires most strongly; furthermore (3) a belief that the agent should perform the action, formed in the absence of any desire to do so, is insufficient to produce a desire to do so in the agent, or even (4) to amplify an existing desire. As I mentioned in the last chapter, I have no use for the philosopher's concept of a desire, which seems to me to be a rather monstrous amalgamation of a preference judgment and an affective attachment. I therefore replace the traditional Humean theory with the following, neo-Humean one: (1) an agent performs an action only if he has an affective attachment to at least one of the expected outcomes of that action, and (2) always acts in a way consistent with a preference for the outcome with the greatest

[14] Representations of affective value are the modern versions of Aristotle's notion of "practical appearances"—*phantasiai*. These are representations of the objects which cause sensations (*aisthêmata*) of pleasure or pain—or more generally, neurological states encoding affective values—as good or bad. Aristotle rightly recognizes that these appearances are shaped by habituation and socialization. See (Reeve 2013, pp. 205, 208). This "tutoring" of one's *phantasiai* is the means by which one transitions from one's first to one's second nature.

affective value; furthermore (3) an agent's consciously forming a belief that he should prefer the expected outcome of one action to that of another—or rather the cognitive neurological processes, obscured from conscious awareness, that are interpreted at the conscious level as the formation of such a belief by a unified self—in the absence of any affective attachment to the expected outcome of the former action, is insufficient to create such an affective attachment, or even (4) to amplify an existing one.[15]

[15] The conscious act of forming a belief, or an intention, is an act of self-interpretation, where what is interpreted is a series of underlying, distributed cognitive neurological processes or events which are obscured from conscious awareness—processes which themselves may admit of the same sort of Bayesian representation applied to arriving at probabilistic judgments at the conscious level. Beliefs themselves have no neurological correlates—there are no "belief-states of the brain." The concept of belief, like the concept of intention and even the concept of action, belongs to what Wilfrid Sellars called the logical space of reasons, not the natural space of causes. The notion of a physically realized, causally efficacious belief-state—which is integral to orthodox philosophy of action—is, to use Sellars' term, a mongrel concept, one which attempts to occupy both spaces at once. The sense of agency—of our beliefs and intentions as the uncaused causes of our actions—results from the neurological processes which produce the symbolic representation of the self, and the interpretation of that symbolic representation. See (Hofstadter 2007). And so in the context of human belief—what we might call "full-blooded belief"—we can say (with Arthur Collins) that to have a belief is to take a stand on some proposition's being true; or (with Robert Brandom) that it is to commit oneself to some proposition's being true. These are normative attitudes with normative implications. But (*pace* Collins) these conscious acts of self-interpretation, which we can refer to as acts of belief-formation, are themselves correlated with neurological processes which take as inputs the outcomes of the underlying neurological processes in need of interpretation. (Though note that if Tyler Burge's anti-individualism is right—and I think it is—we cannot even say that normative attitudes like believing supervene on these neurological processes, since we can individuate acts of believing by the content of what is believed, and this content will depend on a wide range of social, historical, environmental and linguistic features of the world which constitute the context in which the believing agent finds himself.) And these correlate neurological processes have causal implications which constrain future underlying neurological processes and events. A self-interpreting system, like a human being, thereby constrains, through his acts of self-interpretation, the very underlying physical processes he interprets. The self-interpretive, conscious acts of deliberating, believing and intending thus constrain future neurological processes of the brain which produce future actions and form the basis for future self-interpretations. This is how a self-interpreting physical system ends up with self-interpretations, like beliefs, which are bound by the rules which govern the logical space of reasons, through a process of causally constraining the physical processes which are the objects of those self-interpreting acts. Beliefs themselves do not have underlying neurological correlates, even though self-interpreting acts of belief-formation do, precisely because the neurological correlates of these acts are constraint-forming, and so the causally relevant "correlates" of the resulting self-interpretations (beliefs themselves) are *constraints* on other underlying neurological processes—which is to say (following Terrence Deacon), that they are *absences* which determine the space in which other efficient-causal processes occur, the *form* that those processes take. Beliefs in general—whether or not they are occurrent or even formed self-consciously—correspond, from a natural-scientific perspective, to constraints on neurological processes, just as from a normative perspective they are, in one sense, constraints on actions. And that is why, even given a complete causal history of a human agent at the neurological level, we would still need concepts like belief and intention to discern some of the real patterns, in Daniel Dennett's sense of the term, in that agent's past behavior and to predict his future behavior. Those patterns of behavior depend on patterns in the underlying system of causal constraints that emerges, and some of these are only discernible when we see that system as the product of physical processes which

Within the family of Humean theories is a particularly strict version which affirms all four theses, and thus denies that beliefs can have any influence on desire. This is Hume's own view that all reasoning is instrumental—that reason is "the slave of the passions." The neo-Humean correlate of this view would imply that the sort of cognitive neurological processes just alluded to cannot have an effect on affective value states. If this were so, my interpretive framework for representing ends-deliberation at the conscious level would be in tension with the neuroscience of decision-making. But there is some contemporary cognitive neuroscientific research that counts against this strong Humean view. Both fMRI and brain lesion studies have indicated that "cognitive influences descend down to influence the first regions that represent the affective value of stimuli" (Rolls 2009, p. 118).[16] So cognitive processes can interact with representations of affective value, and constrain what affective values expected outcomes are represented as having. I represent such changes in my interpretive framework as updates to the desirabilities of preference-rankings, given updated probability judgments regarding what preference-ranking the agent ought to have, which result in a change of which preference-ranking has the highest desirability, and thus a change in the affective values potential ends are

underlie conscious acts of self-interpretation. So these acts of self-interpretation are far from epiphenomenal or eliminable. That their neurological correlates are correlates of self-interpretations is an essential determinate of their form. Note that this fact offers no relief to those who would defend the existence of state-given reasons with toxin-puzzle arguments. The puzzle must be read either as stating that the correlate neural processes of intending (to which intending is not identical or reducible) will cause the world not to be destroyed, or that the mad scientist will decide not to destroy the world on the basis of his judgment that the agent has the requisite intention (which may or may not be based in whole or in part on an investigation of the agent's neural activity). Whatever the case, the agent's intending is not the (causal) means to the world's being saved, and the agent only has reason to act so as to cause himself to intend to drink the poison, not to intend to do so.

An agent can of course misinterpret himself—in the sense of sincerely asserting (aloud or *sotto voce*) that he believes some proposition *p*, despite his behavior being at odds with actually having taken that stance. We often determine what we believe through the same process used by others, viz., observation and memory of our behavior. At a given point in time, moreover, there may not even be a fact of the matter regarding what someone believes; it may not be possible to determine a specific interpretation of someone's behavior—another person's or one's own—which makes sense of that behavior, until a fuller pattern of behavior emerges in the future. We will return to a number of these issues in Chaps. 13 and 14. As we shall see, we will have to depart from Collins and Brandom insofar as they hold that the possibility of representational thought depends on the prior emergence of normativity and language. The reverse is true, and pre-linguistic creatures can have proto-beliefs, which "ape" the stances taken or commitments made by one with full-blooded beliefs. Our own capacity for believing, and acting in the space of reasons more generally, evolved from this sort of proto-capacity. For an elaboration of some of these ideas, see (Collins 1987; Brandom 1994; Burge 2010; Dennett 1991; van Gelder 1998; Deacon 2011; Gazzaniga 2011). Gazzaniga actually refers to the set of neurological processes correlated with conscious, linguistic acts of self-interpretation as "the interpreter."

[16] This is a specific instance of the general phenomenon described in the previous note, whereby conscious, high-level cognitive processes constrain underlying neurological ones. This finding squares well with Aristotle's contention that *akrasia* can be cured—that the *akratic* agent can start down the path to recovery by mustering the motivation to begin rehabituating himself to find the types of actions he knows to be good pleasant, instead of painful. See (*NE* VII.8 1150b29–35).

represented as having (which I have defined as the values which represent their place in the preference-ranking with the highest desirability). Thesis (4) at least seems to be false.

That still leaves me room to occupy the position of a moderate neo-Humean, who affirms (1)–(3), and it seems to me that my framework, suitably constrained, could be adopted by one who holds such a view. But I myself am willing to conjecture that cognitive processes can result in an increase in the affective value of an option even if that option was previously represented as being indifferent. I therefore reject (3) as well. This might leave me occupying the place of a mild neo-Humean, one who affirms (1) and (2). I do want to commit myself to denying that agents ever act in the absence of an affective attachment to at least one of the expected outcomes of the action performed, in accordance with thesis (1). I reject the opposing position, strong anti-Humeanism. This means I must assume that an actual agent never adopts a deliberative preference-ranking in which outcomes feature as ends when those outcomes are represented as indifferent (or worse) in the ranking with the greatest desirability value. This is a psychological constraint on ends-deliberation that I am willing to accept, so long as we remember that we are referring to the ranking with the highest desirability value *given* deliberation, not prior to it. I also want to allow that an affective attachment may be created or strengthened without there being any corresponding cognitive process or event. So I reject the claim of moderate anti-Humeanism, viz., that though something other than a belief may be required for an agent to be motivated to act (and thus (1) is true), that something else can only be created or amplified by the adoption of the relevant belief. But even mild neo-Humeanism is too strong a position for my view, since I cannot affirm (2). I represent the agent's judgments as having an influence on what first-order preferences he adopts, above and beyond the effect of those judgments on his desirability assignments. This means that I allow for the possibility that an agent will adopt preferences over ends, and act in accordance with those preferences, even though they do not perfectly track the agent's affective attachments to those ends. I do not mean to deny that the preferences of actual agents often do strictly track their affective attachments; only that they must. There are two reasons why they often do. First, many of our ordinary everyday choices—which tie to wear to work, what kind of cheese to put on a sandwich, etc.—are indifferent from the perspective of our judgments about what we ought to prefer; there is no guide other than fancy. Second, when faced with temptation, we may lose our grip on our deliberative preferences and momentarily find them replaced with preferences that strictly track our affective attachments. This is weakness of will, a topic we shall investigate in Chap. 7. So we might describe my specific model of ends-deliberation as being consistent with the truth of *trivial* neo-Humeanism: I merely affirm thesis (1) and deny that affective attachments result exclusively from cognitive processes and events.[17]

[17] Even this statement of trivial Humeanism is too strong in virtue of being overly simplistic. Habitual actions can be performed in the absence of any representation of the affective values of their outcomes, and an agent who plans at time t_0 to perform an action at time t_1 may perform it at

The structure of this model of ends-deliberation fulfills Sen's proposal that we represent an agent's preferences as a synthesis of multiple preference-rankings capturing both the agent's *sympathies* and his *commitments* (Sen 1977, pp. 326–329). The preference-ranking with the greatest desirability value captures the agent's sympathies—his preferences as determined by his own affective attachments and sources of personal satisfaction. This includes both satisfaction that derives from outcomes which directly affect him, and purely sympathetic satisfaction derived from outcomes which only directly affect others.[18] The preference-ranking assigned the highest probability—the ranking which the agent judges most likely to be the one he ought to have—captures the agent's commitments, his judgments about what he ought to pursue.

The structure of the model also reflects the tri-partite division of mental functions into the cognitive, affective and conative, which has provided a general framework for psychological research since the eighteenth century and is currently the object of revived interest (Hilgard 2006). We may accordingly view the model as representing the synthesis of cognitive preferences (the preference-ranking assigned the highest probability) and affective preferences (the preference-ranking assigned the highest desirability) into a single conative preference-ranking. This classification, of course, is simply an appropriation of the Platonic/Aristotelian view of the functions of the soul.[19] None of this is to say that we must or should assume that traditional tri-partite functional psychology is an accurate scientific theory of the mind. But the lasting usefulness of this way of thinking as an interpretive framework for representing human deliberation and action is something to be marveled at. I suspect, moreover, that one of the major sources of strongly Humean sympathies nowadays is a tendency to interpret preferences over outcomes in decision-theoretic models of instrumental deliberation as affective preferences. The cognitive function is then limited to making judgments about the probability that such-and-such

---

t_1 even though when t_1 arrives the outcome is not represented as having any affective value. Habitual actions, however, are borderline cases of action, and in any event it is unlikely that a habit could be successfully cultivated by an agent if the agent had no affective attachment to the outcome of the action at the time he was cultivating it. Ceasing to have an affective attachment to an outcome should trigger a revision of one's intentionally adopted preferences, according to my model. It is interesting that for actual agents this may fail to happen (and even if it does happen, an actual agent may forget that he has changed his mind, and proceed with the earlier plan when the time for action arrives). But it is again unlikely that the original plan would have an effect on the agent's behavior if the agent did not have an affective attachment to the outcome at the time the plan was adopted. The relationship between affective value, motivation, and action is thus more complicated than trivial Humeanism makes it out to be; but these complications need not trouble us unduly so long as we are aware of them. For more details on the neuroscience pertaining to these issues, see (Schroeder 2004). Note that Schroeder identifies affective value with the philosopher's concept of a desire, a position which I do not find to be defensible.

[18] We could derive a narrowly egoistic preference-ranking by determining what the agent's desirability assignments would be if he believed that others were indifferent to all outcomes.

[19] The fit with Aristotle's theory of the varieties of desire is closer. In Aristotelian terms, the cognitive preferences reflect wish, the affective preferences reflect appetite, and these combine to determine the conative preferences, which reflect the impulse (ὄρεξις) that initiates action. See (*Metaphysics* XII.7 1072a27–28 cited in Reeve 2013, p. 206). For an interesting recent study of Aristotle's views on the functions of the soul, see (Johansen 2012).

an outcome will result from such-and-such an action. There are then no cognitive preferences at all; the agent's preferences over actions, determined in accordance with the expected utility principle, are conative. But for the agent who has engaged in ends-deliberation, those preferences over outcomes with which instrumental deliberation begins are no longer merely affective. They are themselves deliberately and intentionally adopted following a synthesis between the agent's affective and cognitive preferences over ends.

3.2 Forms of Ends-Deliberation

Once the agent has a deliberative ranking of ends, the process of ends-deliberation can take on a variety of forms, each of which is designed to solve a type of practical problem that requires selecting ends. The three main varieties of ends-deliberation after the initial stage are (1) deliberating about promoting ends; (2) deliberating about adopting new ends; and (3) deliberating between specifications of ends. I will discuss each in turn.

Deliberation that results in the promotion of an end is the easiest to model. It is essentially the same process as the initial round of ends-deliberation. An agent promotes an end within his preference-ranking on the basis of evidence that something he is pursuing is more choiceworthy than he has been taking it to be. Recall that we have defined ends purely in terms of the preference relations between outcomes in the agent's preference-ranking. These relations determine whether one end is taken to be more, less, or as valuable as another, and whether one end is sought for the sake of another. The process of promoting ends, then, is just the process of changing which preference-ranking one has adopted. This is done *via* the method of ends-deliberation described above. The agent in this situation is deciding between *continuing* to endorse the preferences he has deliberately adopted, and adopting new preferences.

The second form of ends-deliberation is deliberation about a new end, the existence of which the agent was previously unaware. We represent the agent as becoming aware of a new end by replacing each preference-ranking R_i in the agent's meta-preference with a set of preference-rankings R_i^+. Each preference-ranking in this set will be identical to the old ranking R_i, except that the new rankings will each incorporate the new end into one of the values (in Ramsey's sense) already appearing in the ranking, such that the new end appears once in every value. A particular first-order preference-ranking may now be denoted $R_{i,\nu}$. This will be the expansion of the old preference-ranking R_i, with the new end placed in value ν within that ranking. The new end will have a perfectly well-defined valuation within that preference-ranking—it will have the valuation that attaches to whatever value (again in Ramsey's sense) it is placed in. We denote the agent's new, expanded meta-preference as R^+; this replaces the agent's old meta-preference ranking R.

We are able to retain the assumption that the original preference-ranking R_i was complete, even though it did not include the new end. First, new ends are either ends which did not exist previously, or ends of which the agent was utterly unaware. There were no ends of which the agent was aware that were left unranked in R_i. The

instant the agent becomes aware of a new end, he incorporates it into his preferences. That is to say, he *replaces* his old preference-ranking with a new one that includes the new end—the new preference-ranking being one that appears in his new, expanded meta-preference-ranking. We will see how he does this in a moment, and that it is done in a way that maximizes psychological continuity, so that the agent can truly be said to have the *same* preferences *except* for the fact that the new end has been incorporated. So the agent always has a complete preference-ranking over all ends of which he is aware. Second, the old preference-ranking (and all the other possible preference-rankings) is complete at the level of values. Ramsey's axioms guarantee that the old preference-ranking will be dense—that is, that between any two values in it, a third value is found. We further assume that the agent's initial preference-ranking is also sufficient, in the sense that the agent will not judge any newly discovered end better (or worse) than he had ever judged anything to be. Any new end that appears in an expanded preference-ranking, therefore, must appear in a value that was already present in the original preference-ranking. There is simply nowhere else it could end up. This is yet another advantage, from the perspective of modeling ends-deliberation, of working with a background decision-theory that has the structure of Ramsey's.

We can then represent the agent's choice to adopt one extended preference-ranking rather than another—and thus his choice about how to incorporate a new potential end into his preferences—as follows:

$$i\left(R_{i,\nu}\right) \geq i\left(R_{j,\mu}\right) \leftrightarrow des_{new}\left(R_{i,\nu}\right)\sum_{a=1}^{k} p\left(R_{i,\nu} \mid E_a\right) p_{new}\left(E_a\right) \geq des_{new}\left(R_{j,\mu}\right)\sum_{a=1}^{k} p\left(R_{j,\mu} \mid E_a\right) p_{new}\left(E_a\right)$$

with the agent adopting the ranking that satisfied: max $des(R_{x,\chi})p_{new}(R_{x,\chi})$. The valuations of $des(R_{i,\nu})$ and $des(R_{j,\mu})$ are determined according to the two rules for desirability updating given above, with $des_{old}(R_{i,\nu})$ set equal to $des_{old}(R_i)$ for all ν and $des_{old}(R_{j,\mu})$ set equal to $des_{old}(R_j)$ for all μ—i.e. we represent the agent as being initially indifferent to the location of the new end in his preference-ranking and as initially valuing any ranking that contains it identically to the corresponding ranking before he became aware of the new end. This is our primary way of preserving psychological continuity. He then updates those desirability values after he has acquired evidence bearing on his judgments of the expanded rankings. $p_{old}(R_i^+)$—i.e. the probability that one or another of the rankings in R_i^+ is correct—is initially set equal to $p_{old}(R_i)$, with all rankings in R_i^+ initially equiprobable. We must assume that when an agent deliberates about how to incorporate a new potential end into his preferences, he holds all of his other preferences fixed (a point which also helps to preserve psychological continuity). Call R_1 the agent's top-ranked preference-ranking—the one he currently adopts—before he becomes aware of the new potential end. Then $R_{1,\nu}$ and $R_{1,\mu}$ are two different ways of expanding R_1 so as to include the new end in one of R_1's values. The new, expanded preference-ranking actually adopted by the agent would then be determined as follows:

$$i\left(R_{1n}\right) \geq i\left(R_{1m}\right) \leftrightarrow des_{new}\left(R_{1v}\right)\sum_{a=1}^{k} p\left(R_{1v} \mid E_a\right) p_{new}\left(E_a\right) \geq des_{new}\left(R_{1\mu}\right)\sum_{a=1}^{k} p\left(R_{1\mu} \mid E_a\right) p_{new}\left(E_a\right)$$

with the agent adopting the ranking that satisfied: max $des(R_{1,\chi})p_{new}(R_{1,\chi})$.[20]

The third variety of ends-deliberation is deliberating about alternate specifications of ends. Let us begin with another one of Richardson's examples (Richardson 1986, pp. 171–173). An environmentalist is trying to decide between two products. He is struggling, because both seem equally likely to enable him to contribute to protecting the environment, and thus he should be indifferent between them. But he does not feel indifferent. The first step toward resolving this problem is to specify his guiding norm ("protect and preserve the environment"), in different ways, each of which supports one of his two possible decisions but not the other. He engages in some specificational reasoning, and formulates two specifications: (1) preserve the as-yet untouched portions of the wilderness; and (2) protect the integrity of urban environments by minimizing pollution. He observes that the first norm supports buying the first product, while the second supports buying the second. The agent must now select one of these two specifications. Richardson's theory is of little help to us here—it simply tells us that the agent should realize which of these two ends better coheres with his other ends (Richardson 1986, p. 173). My theory, however, has more to offer.

What the agent has realized by confronting this choice is that where he thought there was a single end occupying one value of his preference-ranking, there have turned out to be two, and each of his two available actions will enable him to pursue one. He engages in specificational reasoning to make the distinction between the two clear. His new awareness of the distinction is then modeled by replacing his meta-preference-ranking with a new one in which both specified ends are represented in each possible preference-ranking. By default, he is initially indifferent between the two newly specified ends, and values both as he did the unspecified end—i.e. his old preference-ranking is replaced by a new one which is identical to it except for the fact that the two specified ends are now found in place of the unspecified one. But he suspects that additional evidence would lead him to prefer one to the other. He remains indifferent between them as he acquires the additional evidence. He will then update his judgments about them, and then update his desirabilities in accordance with the two rules for desirability updating given above. On

[20] It would be desirable to model the agent's growing awareness of new ends in a smoother and more continuous way, rather than with the abrupt replacement of one meta-preference ranking by another that is used here, if it were possible to do so. But it may not be possible. Edi Karni and Marie-Louise Vierø have succeeded in doing just this, in the case of an instrumental reasoner whose awareness of new material consequences is growing. However, by "material consequences", they mean outcomes which can be transformed by the agent into the satisfaction of basic preferences; and their model relies on the assumption that the agent's basic preferences are permanently fixed (Karni and Vierø 2013). For an excellent overview of the rapidly expanding literature on awareness in formal epistemology, see (Schipper 2015).

the basis of these updates, he will reposition the two specified ends in his preference-ranking. His subsequent choices of actions will reflect this.

3.3 Modeling Deliberation about Ends as a Dynamic System

I have shown that an agent's degree of credence in a preference-ranking's being the one he ought to adopt can be represented by a probability function, and that his degree of attachment to the possibility of a preference-ranking's being the one he ought to adopt can be represented by a desirability function. And I have argued that an agent's meta-preference-ranking should be determined by his degrees of credence in and attachment to the preference-rankings appearing in the meta-ranking. The model developed thus far includes update rules for both probabilities and desirabilities. It is, moreover, an endogenous model of evidence-based preference formation, as it explicitly models the agent's formation of beliefs about what counts as a reason for having one preference rather than another, and how weighty those reasons are, in terms of the agent's experiences and observations of the world. And it is a model of the process of deliberating about what preferences to adopt. But it is not yet a truly *dynamic* model of the process of deliberating about ends. The model developed thus far represents agents as immediately recalibrating their preferences in response to new evidence. In order to construct a genuinely dynamic model, we must represent agents as sometimes occupying genuine states of indecision—as having abandoned (some of) their current preferences, and yet as being too uncertain about what new preference-ranking they ought to adopt to make a commitment to a new one. A dynamic model of rational deliberation about ends will then model the process of reasoning which brings the agent from such a state of indecision to a state of deciding to adopt a particular preference-ranking—it will be a model of the agent's process of *figuring out and making up his mind about what to value*.

In order to model this process dynamically, it is absolutely essential that we have chosen a Ramsey-style theory as our static decision-theoretic background theory. I emphasize again that we read the biconditional—that the agent will prefer action *a* to action *b IFF* the expected value of *a* is greater than the expected value of *b*—as a rational requirement in addition to the axioms of the theory: that is the preference over actions the agent should have, and the one he will have, if he is rational, given his probability judgments and his preferences over outcomes. This contrasts with versions of decision theory which start with an axiomatically restricted set of preferences over actions. In such theories, the goal is to demonstrate that the axioms restricting the set of preferences over actions are themselves both necessary and sufficient for a representation of those preferences in terms of the expected value of the actions. The biconditional then expresses a necessary and provable connection between the agent's preferences over actions and the expected value of those actions. What this means is that in theories of this type, it is impossible to break up the biconditional, such that we continue to assert:

$$i\left(R_1\right) \geq i\left(R_2\right) \rightarrow des_{new}\left(R_1\right) p_{new}\left(R_1\right) \geq des_{new}\left(R_2\right) p_{new}\left(R_2\right);$$

but do not assert the converse:

$$des_{new}\left(R_1\right)p_{new}\left(R_1\right) \geq des_{new}\left(R_2\right)p_{new}\left(R_2\right) \rightarrow i\left(R_1\right) \geq i\left(R_2\right).$$

Up until this point, we have been accepting both of these conditionals. But from here on out, we only accept the former. We will, as our background decision theory allows us to, break up the biconditional, and maintain that an agent can assign a highest expected value to one preference-ranking without the consequence of adopting that preference-ranking. Instead of this simple claim, we will put in place a more complex dynamic mechanism, which will specify the more elaborate conditions under which a rational agent adopts a preference-ranking which has a greater expected value than the other preference-rankings he might adopt.

I have taken my description of a state of indecision from the work of Brian Skyrms. Skyrms has developed a dynamic model for instrumental reasoning, which contains all the essential structural elements needed for a dynamic model of ends-deliberation. I will therefore provide a brief discussion of Skyrms' model, and then turn to the construction of a dynamic model of ends-deliberation in a way that follows that lead.

3.3.1 Skyrms' dynamic model of instrumental deliberation

Skyrms has two guiding commitments. First, that the principle of expected value is the touchstone of a theory of practical rationality; and second, that we should understand practical deliberation as procedural (Skyrms 1990, p. 2). Skyrms thus describes an agent who is trying to figure out what to do as occupying a state of indecision. In such a state, the agent has assigned expected values to his available actions, based on the value to him of the possible outcomes of those actions and his judgments of the probabilities of those outcomes being realized. And it may be that there is one action which currently has a higher expected value than any other. But the agent continues to occupy a state of indecision because he is not yet sufficiently confident that that action is the one which will end up having the highest expected value, and so he is not yet prepared to commit to it. He is aware that he may receive new information which will lead him to revise his probability judgments in a way that has a significant impact on the expected values of the actions he is considering.[21] Remaining in this state of indecision is the agent's *status quo*, which we may denote q. There is a value—$V(q)$—associated with remaining in the status quo. This is just the average of the expected utilities of the available actions. Thus, in Skyrms' model, although actions are evaluated according to their expected value, the fact that an

[21] There is a superficial similarity between Skyrms' model and Kreps' "preference for flexibility." But the reader should keep in mind that Kreps has developed a static model for representing the current preferences of a rational agent who is aware of the fact that his tastes may be different in the future, when the time for decision arrives. Skyrms' has, by contrast, developed a dynamic model of an agent who is in the process of reaching a conclusion about what action he ought to perform, and is not yet satisfied that he has incorporated enough of the relevant information to arrive at such a conclusion.

action has an expected value greater than that of any other action does not in itself imply that the agent prefers that action—the agent may, instead, prefer to remain in the status quo for the time being, and hold off on forming a preference for any action. A dynamic model of deliberation thus requires the analysis given above of the relation between expected value and preference over actions. An agent in a state of indecision assigns a probability to the *eventual* performance of each of his available actions.

Agents move from one state of indecision to another, with each step bringing them closer to making a decision, in accordance with a rule which determines the way in which these probabilities on eventual performance are revised. Skyrms suggests that any rule for revising the status quo should *seek the good*, in that:

(i) It raises the probability of the eventual performance of an action only if that act currently has an expected value higher than that of the status quo.
(ii) It raises the sum of the probabilities of all actions with expected values higher than that of the status quo.

One such (intuitively appealing) rule has been given by Nash. Define the *covetability* of an action in a state of indecision as (1) the difference between the expected value of the action if the action is performed and the expected value of the status quo if the former is greater than the latter; and (2) 0 otherwise. So for an action A we have:

$$Cov(A) = \max\left[V(A) - V(q), 0\right].$$

The rule is then the *Nash map*, which takes the agent from state of indecision q to a new state of indecision q' by changing the probability p_i of the eventual performance of each action A_i as follows:

$$p_i' = [p_i + Cov(A_i)] / [1 + \sum_i Cov(A_i)].$$

In continuous time, we have the corresponding *Nash flow*:

$$dp(A)/dt = \left[cov(A) - p(A) \bullet \sum_i cov(A_i)\right] / \left[1 + \sum_i cov(A_i)\right].$$

This describes the change in the probability assigned to the eventual performance of the actions in continuous time.

The centerpiece of Skyrms' model is the thought that the process of deliberation itself—the process of revising one's judgments regarding the probabilities that one will eventually perform various actions—can itself generate new information relevant to the agent's decision, information that affects the agent's probability judgments regarding states of the world (Skyrms 1990, p. 2). "[P]robabilities," he asserts, "can change as a result of pure thought" (Skyrms 1990, p. 2). And "where deliberation generates new information relevant to the decision under consideration, a rational decision maker will (costs permitting) feedback that information and reconsider" (Skyrms 1990, p. 1). Let us see, then, how Skyrms builds this thought into his model of deliberation.

Skyrms defines the *personal state* of an agent as the state determined by the agent's state of indecision (including the probabilities assigned to the eventual performance of the various actions) and the agent's state of nature (including the probabilities assigned to the obtaining of the various possible states the agent might occupy). The agent's *personal state space* is then the product space of the space of indecision (all the states of indecision it is possible for the agent to occupy) and the space of states of nature (all the states of nature, with all the possible probability distributions over them, which is it possible for the agent to occupy). Let X be the space of indecision with $x \in X$ a particular state of indecision; and let Y be the space of states of nature with $y \in Y$ a particular state of nature. Then the agent's personal state space is $X \times Y$, and the agent occupies a particular personal state $<x,y>$.

Deliberation defines a dynamics on the agent's personal state space. This means that the total process of deliberation is modeled as a dynamical function ϕ which maps the agent's personal state $<x,y>$ onto a new personal state $<x',y'>$. So we have ϕ: $X \times Y \rightarrow X \times Y$, and $\phi<x,y> = <x',y'>$. The function ϕ has two associated rules:

(i) What Skyrms calls the "adaptive dynamical rule," and denotes D. This is any rule which seeks the good, of which the Nash map is an example. This rule maps $<x,y>$ onto x'.
(ii) The informational feedback process, denoted I, which maps $<x,y>$ onto y'.

Rule I revises the agent's probability judgments regarding the state of nature, as determined by the new information generated by the application of rule D. Since those revisions alter the expected value of the available actions, they alter the expected value of the status quo and the covetability of each action as well. A change in covetability results in another round of revisions to the probabilities of eventual performance of the actions, in accordance with rule D. Since this leads to another application of rule I, we have a feedback loop.

This process will continue until the agent arrives at a *deliberational equilibrium*, at which $\phi<x,y> = <x,y>$. Skyrms' main result is that if D is a rule that seeks the good and $I : Y \rightarrow Y$ is continuous, then it follows from Brouwer's fixed-point theorem that a deliberational equilibrium $<x_e,y_e>$ exists for $\phi = <D,I>$. Once the agent reaches a deliberational equilibrium and the feedback process stops, the agent no longer has doubts regarding the stability of the expected value of the actions sufficient to keep him in a state of indecision. He thus commits to a preference for the action with the highest expected value at this point.

Skyrms' real achievement, however, is not that he has devised a way to model instrumental deliberation dynamically—as important as that is. It is that by modeling instrumental deliberation in this way, he is able to *embed decision theory in game theory*. Assume that two or more Bayesian deliberators, each of which uses a rule that seeks the good, are engaged in a non-zero-sum non-cooperative game—where a game is simply a strategic interaction of some kind. Each player in the game knows that her decision about what action to take, together with the decisions of the other players, will determine which outcome each player in the game attains. Assume further that each player's good-seeking rule is known to all the others, as is

each player's initial state of indecision, and it is common knowledge that these are known by all.

With these assumptions in place, each player is able not only to move from one state of indecision to another in accordance with his own rule; he also knows that all the other players will be doing the same, and has all the information he needs in order to perform the same calculations they perform to determine what their new states of indecisions will be. Skyrms calls this "updating by emulation." Each player is able to do this, and that fact is common knowledge. We already know that each of these Bayesian agents individually will eventually reach a deliberational equilibrium. Skyrms shows that, in addition, all the players will be at a deliberational equilibrium if, and only if, they have arrived at a *Nash equilibrium* for the game. A Nash equilibrium is a combination of players' strategies, each one of which counts as a *best response* to the strategies of all the other players. A player's strategy is a best response just in case, assuming the players all know each other's strategies, the strategy maximizes the player's expected value.

3.3.2 A dynamic model of rational deliberation about ends

Skyrms' dynamic model of instrumental deliberation can be translated into a dynamic model of ends-deliberation in a fairly straightforward way. We assume the agent has moved from his pre-deliberative attachments to a state of indecision regarding what preference ranking he should adopt, having been confronted with facts that he takes to be evidence against holding onto his initial attachments. So the agent has probabilities over the eventual adoption of various preference-rankings, the adoption of each of which has an associated expected value. This expected value is determined by the agent's degree of credence in the correctness of the preference-ranking ($p(R_i)$) and degree of attachment to it ($des(R_i)$), and is computed in the way already described. The space of indecision, the space of states of nature, and the agent's personal state space are all defined as above. The agent moves from one state of indecision to another by first determining the covetability of adopting a given preference-ranking (here, for notational simplicity, I use R_i to denote the act of intentionally adopting preference ranking R_i):

$$Cov\left(R_i\right)=\max\left[V\left(R_i\right)-V\left(q\right),0\right].$$

and then revises his probability judgments regarding eventual adoption according to the Nash map:

$$p_i'=[p_i+\left[Cov\left(R_i\right)\right]/[1+\sum_i Cov\left(R_i\right)].$$

This deliberative act, in which the agent revises his preference judgments regarding eventual preference adoption, can itself generate new information relevant to the agent's probability judgments regarding the correctness of the preference-rankings he is trying to decide among. In particular, the act of revising one's probability judg-

ments regarding eventual performance can elicit a heretofore hidden emotional response to the idea of adopting one preference-ranking or another. This emotional response is an observed fact which constitutes evidence in the light of which one's probability judgments regarding the correctness of the various preference-rankings will be updated. The rule *I* is then the probability update rule that determines the revision of those judgments. This is a precise, formal way of explicitly modeling the moments of self-realization that play such a large (but unanalyzed and unexplained) role in the vignettes of Richardson and Millgram. The deliberative acts of those agents themselves generate new information which is relevant to their views on the possible preferences among which they are deliberating.

At this stage in the process of ends-deliberation, we must introduce an element not present in Skyrms' instrumental model. A change in one's degree of credence in the correctness of a preference-ranking can result in a change in the desirability one attaches to discovering that that is the preference-ranking one should adopt, in accordance with the rule given above:

$$des_{new}\left(R_n\right) = des\left[R_n \mid p_{new}\left(R_n\right)\right].$$

So deliberating about ends can lead to a change in the expected value of the objects of deliberation which is determined by both revisions of probabilities and the revisions of desirabilities which they lead to. Since the revision in desirability cannot in turn lead to another revision in the probability, there is no feedback loop at this stage.

With the expected value of adopting a given preference-ranking thus revised, the covetability of adopting it, and the expected value of the status quo, will of course change. And thus the probabilities of eventual adoption will again be revised. That revision initiates another round of the feedback loop between *D* and *I*. Skyrms' result securing the existence of a deliberational equilibrium from Brouwer's fixed-point theorem carries over, and so at some point the dynamical deliberation function, which we may denote ψ to distinguish it from its instrumental cousin, will yield $\psi < x,y > = < x,y >$, and the process of deliberation will halt. At this point, the agent will be ready to commit to adopting a new preference ranking.

Of course, we expect agents to continue to revise and refine their preferences as they experience more of life and the world, as their tastes change, and as they age. A new experience which triggers a revision in the probabilities or desirabilities assigned to the agent's possible preference rankings can occur at any time, in accordance with the update rules given above:

$$p_{new}\left(R_n\right) = \sum_{i=1}^{k} p\left(R_n \mid E_i\right) \cdot p_{new}\left(E_i\right),$$

and

$$des_{new}\left(R_n\right) = des\left(R_n\right) + \sum_{i}\left[des_{new}\left(E_i\right) - des\left(E_i\right)\right] \cdot p\left(E_i \mid R_n\right).$$

The experiences that trigger these revisions are the model's exogenous shocks, and the model takes their occurrence as placing the agent back in a state of indecision. By leading to a change in $V(R_n)$ they lead to a revision of the covetability of adopting the preference-ranking, and thus to a revision in the probability of eventually deciding to continue to embrace a given preference-ranking, or to adopt a new one. Another way to get thrown back into a state of indecision is to become aware of a new end or realize that one of one's ends is in need of specification. In these cases one's entire meta-preference is replaced by an expanded one, and the state of indecision will feature new expanded preference-rankings from that new meta-preference, one of which one will end up adopting.

So in place of the assertion that $des_{new}(R_1)\,p_{new}(R_1) \geq des_{new}(R_2)\,p_{new}(R_2) \rightarrow i(R_1) \geq i(R_2)$,

we will now assert that a rational agent will prefer to intentionally adopt a preference ranking R_1 rather than a preference-ranking R_2 ($i(R_1) \succ i(R_2)$) only if:

(i) the agent judges that $des_{new}(R_1)p_{new}(R_1) \geq des_{new}(R_2)p_{new}(R_2)$;
(ii) the agent has reached a deliberational equilibrium $\psi < x,y > = < x,y >$.

In place of (ED), we now have:

(ED*): (1) If a rational agent prefers to intentionally adopt a preference-ranking R_1 rather than a preference-ranking R_2, ($i(R_1) \succeq i(R_2)$) then condition (i) above is satisfied; and (2) a rational agent will only prefer to adopt R_1 rather than R_2 if conditions (i) and (ii) above are satisfied.

With (ED*), we finally have a decision rule for the intentional adoption of preferences over ends by a rational agent.

Agents do not deliberate about and adopt ends in a vacuum. They do so as members of a society, interacting in various ways with others. The decisions made by others about what ends they will adopt affect one's own decisions about what ends to adopt, and may partially determine which ends one ought to adopt. It is easy to see why this should be so. As we have seen, one of the major types of evidence that one should adopt a particular preference-ranking over ends is that one can achieve many of its high-ranking ends without failing (at least most of the time) to perform actions that advance the interests of others, and without failing to perform actions that one has categorical reasons to perform. Which actions satisfy these criteria depends to some significant extent on which ends other agents select for themselves. So deliberation about ends must ultimately be modeled as a type of strategic interaction. And a non-zero-sum non-cooperative game is precisely the correct type. Others' choices of ends, *via* the reasons grounded by these choices, affect the probability that a given preference-ranking is the ranking one ought to adopt for oneself. This in turn affects the expected value, as I have defined this notion in the context of ends-deliberation, of choosing a given preference-ranking over ends. So the expected value of a given choice depends not only on what choice one makes, but also on what choices the other agents make. There might be some preference-ranking which would be the one a given agent ought to adopt, *if only the rest of humanity would cooperate and go along with one's wishes*. This would be the ranking the agent wishes were the one he ought to adopt. But each agent has his own life

to lead, and none will expect all others to choose precisely those ends which will most effectively limit his duties to others, or which are most conducive to his own aspirations. So we each attempt to choose the best set of ends we can, given what we know about others' attempts to do the same, and aware of the normative ties that bind us to each other.

Note that it is only individuals' *selection* of ends that has been modeled as a non-cooperative game. Efforts to *achieve* the ends selected may very well be cooperative. For one, agents may end up choosing ends whose achievement requires cooperation, and decide to cooperate in order to achieve them. But more fundamentally, the fact that the reasons grounded by the interests and choices of others are modeled as having an effect on the expected value of the choices of each individual agent shows that the agents are selecting ends with the knowledge and intention that, after the ends have been selected, they will be cooperating in the course of trying to achieve their ends in a more basic way. Even if no specific ends are shared among agents, each agent will pursue his own ends in ways that respect the normative ties that exist between him and other agents. Each agent will act in a way that respects and responds to the reasons grounded by the interests and choices of others in the course of pursuing his own ends.

Each member of a group of Bayesian agents, engaged in the game of selecting ends of the kind just described, will arrive at a deliberation equilibrium—and thus select a preference-ranking over ends—if, and only if, taken together, their choices of what preferences to adopt constitute a Nash equilibrium for the game. So the stable position which an individual agent reaches in arriving at a deliberational equilibrium is replicated at the social level of interacting agents. This completes our development of a formal, dynamic, endogenous model of rational deliberation about ends.

4 Conclusion: On Authenticity

With my account of the competence aspect of autonomy complete, I would like to say a few words about the authenticity aspect of autonomy, and thus complete an account of the rational dimension of the capacity of autonomy. My theory allows for a more precise characterization of the authenticity component than has so far been offered. Dworkin and Frankfurt characterize authentic agents as ones who make reflective second-order endorsements of their first-order preferences. According to my theory, authenticity is achieved when an agent who prefers one ranking to another—$i(R_1) \succ i(R_2)$—makes that preference judgment on the basis of a judgment that that preference more likely than not reflects what is actually valuable according to the available evidence, $p(R_1) > p(R_2)$, *and* is satisfied at finding this to be the case, $des(R_1) > des(R_2)$. An authentic agent has achieved an alignment, not between his first-order preferences and his second-order preferences, but between all of (i) his first-order deliberative preferences over acts of intentionally adopting preference-rankings over ends; (ii) his second-order attachment-based preferences over those

preference-rankings over ends; and (iii) his probability judgments about which preference ranking over ends he ought to adopt. Authenticity, in short, is harmony among an agent's affective attachments, judgments about what is valuable, and choices.[22] It is harmony between the functions of his mind, the parts of his soul.[23] The potential for achieving authenticity thus rests on developing competence in exercising the rational capacity for ends-deliberation. Even after the agent constructs a deliberative preference-ranking over the actions of intentionally adopting his various possible preference-rankings over ends, his meta-preference-ranking over those rankings—which expresses his attachment to them—remains. The two are distinct. And they may of course be out of sync with one another—it may be that $i(R_1) \succ i(R_2)$ even though $des(R_1) < des(R_2)$. Alternatively, it may be that $i(R_1) \succ i(R_2)$ and $des(R_1) > des(R_2)$ but $p(R_1) < p(R_2)$. So even after the agent has deliberatively and intentionally adopted preferences over ends, the goal of realizing the other aspect of autonomy—authenticity—may remain. Progress toward this goal is possible by seeking additional evidence to strengthen judgments (and trigger further updates to desirability), and by cultivating his taste for reasons.

Since for the excellent ends-deliberator, desirabilities are in line with judgments, the fullest exercise of the competence aspect of autonomy coincides with the achievement of authenticity. These are two aspects of the single, rational dimension of autonomy.

References

Aristotle. *De Anima.* Oxford classical texts, ed. W.D. Ross. Oxford: Oxford University Press.

Aristotle. *Metaphysics* [*Metaphysica*]. Oxford classical texts, ed. W Jaeger. Oxford: Oxford University Press.

Aristotle. *Nicomachean Ethics* [*Ethica Nicomachea*]. Oxford classical texts, ed. L Bywater. Oxford: Oxford University Press.

Aristotle. *Prior and Posterior Analytics* [*Analytica Priora et Posteriora*]. Oxford classical texts, ed. W.D. Ross. Oxford: Oxford University Press.

Aristotle. *Topics and Sophistical Refutations* [*Topica et Sophistici Elechi*]. Oxford classical texts, ed. W.D. Ross. Oxford: Oxford University Press.

Barnes, J. 1994. *Aristotle's posterior analytics*. Oxford: Clarendon Press.

[22] We can translate the view of authenticity that falls out of my theory into Frankfurt-Dworkin vocabulary fairly easily. By "first-order preferences," they mean the particular preference-ranking which sits atop the agent's attachment-based meta-ranking. By "second-order preferences," they mean something much closer to what I have represented as probability judgments regarding the correctness of preference-rankings. So "bringing first-order preferences in line with second-order ones" becomes, in my model, cultivating revisions in desirabilities such that the preference-ranking with the highest probability value also has the highest desirability value. And of course, when that is the case, that preference-ranking will be the one intentionally adopted. My account of authenticity is thus a more precise version of theirs.

[23] This is one of Aristotle's ways of describing the good man (*NE* IX.4 1166a12–13). We will see in Chap. 14 that the *phronimos*, on my construal as the excellent deliberative agent, is also the authentic agent *par excellence*.

Bradley, R. 2008. Preference kinematics. In *Preference change*, ed. T. Grune-Yanoff and S.O. Hansson, 221–242. Dordrecht: Springer.
Brandom, R. 1994. *Making it explicit*. Cambridge: Harvard University Press.
Burge, T. 2010. *Origins of objectivity*. Oxford: Oxford University Press.
Collins, A. 1987. *The nature of mental things*. South Bend: University of Notre Dame Press.
Cooper, J. 1986. *Reason and human good in Aristotle*. Indianapolis: Hackett.
Deacon, T. 2011. *Incomplete nature: How mind emerged from matter*. New York: WW Norton.
Dennett, D. 1991. Real patterns. *Journal of Philosophy* 88(1): 27–51.
Frede, D. 1995. The cognitive role of *phantasia* in Aristotle. In *Essays on Aristotle's De Anima*, ed. A. Oksenberg-Rorty and M. Nussbaum, 280–296. Oxford: Clarendon Press.
Gazzaniga, M.S. 2011. *Who's in charge? Free will and the science of the brain*. New York: Harper Collins.
Hilgard, E.R. 2006. The trilogy of mind: cognition, affection, and conation. *Journal of the History of the Behavioral Sciences* 16(2): 107–117.
Hofstadter, D. 2007. *I am a strange loop*. New York: Basic Books.
Irwin, T.H. 1987. *Aristotle's first principles*. Oxford: Clarendon Press.
Jeffrey, R. 2004. *Subjective probability: The real thing*. New York: Cambridge University Press.
Johansen, T.K. 2012. *The powers of Aristotle's soul*. Oxford: Oxford University Press.
Karni, E., and M.-L. Vierø. 2013. 'Reverse Bayesianism': A choice-based theory of growing awareness. *American Economic Review* 103(7): 2790–2810.
Kolnai, A. 1961. Deliberation is of ends. *Proceedings of the Aristotelian Society New Series* 62: 195–218.
MacIntyre, A. 1985. *After virtue*. South Bend: Notre Dame University Press.
Nozick, R. 1989. *The examined life*. New York: Simon & Schuster.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Reeve, C.D.C. 2013. *Aristotle on practical wisdom: Nicomachean ethics VI, translated with an introduction, analysis, and commentary*. Cambridge, MA: Harvard University Press.
Richardson, H. 1986. *Practical reasoning about final ends*. Cambridge: Cambridge University Press.
Rolls, E.T. 2005. *Emotion explained*. Oxford: Oxford University Press.
Rolls, E.T. 2009. From reward-value to decision making: Neuronal and computational principles. In *Handbook of reward and decision making*, ed. J.-C. Dreher and L. Tremblay, 97–133. New York: Elsevier Academic Press.
Schipper, B.C. 2015. Awareness. In *Handbook of epistemic logic*, ed. H. van Ditmarsch, J.Y. Halpern, W. van der Hoek, and B. Kooi. London: College Publications.
Schroeder, T. 2004. *The three faces of desire*. Oxford: Oxford University Press.
Sen, A. 1977. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy and Public Affairs* 6(4): 317–344.
Skyrms, B. 1990. *The dynamics of practical deliberation*. Cambridge, MA: Harvard University Press.
Stocker, M. 1990. *Plural and conflicting values*. Cambridge: Cambridge University Press.
van Gelder, T. 1998. Monism, dualism, pluralism. *Mind and Language* 13(1): 76–97.

Part II
Liberty: Freedom

Chapter 5
Liberty: Freedom – Introduction

We now have at our disposal a far more robust and precise conception of individual autonomy than has ever been developed. In the next two chapters I will articulate a similarly robust and precise conception of individual freedom. I will conclude Chap. 7 by defining a compound conception of individual liberty, the guiding value of political liberalism, which unites my accounts of autonomy and freedom. The account of individual liberty which I propose at the end of these two chapters will have the singular virtue of being fully measurable (at least from a theoretical perspective, though a good deal of work remains from an operational one). For each component of liberty as I shall define it, it will be possible to compare the extent of liberty enjoyed by any two agents. Once we have a measurable conception of liberty, we will be able to speak meaningfully of the distribution of liberty among the members of a society. I will then argue, in Division II of the book, for an egalitarian theory of distributive justice that takes liberty to be the appropriate object of the State's distributive concern.

One of the questions which will command our attention in Chap. 9 is that of why we should take liberty rather than welfare or well-being (however one may care to interpret these latter two notions) to be the appropriate object of the State's distributive concern. While the full answer will have to wait until the point in the discussion when the question itself naturally emerges, I should declare at the outset that the reason is not that well-being is a matter of little importance according to my view. Far from it. The conception of liberty which I am developing here in Division I is unintelligible apart from a view of what constitutes well-being—for as we shall see, one of the ways in which my conception of liberty may be fairly characterized is as the liberty to pursue and achieve well-being. In this introductory chapter, therefore, I will briefly discuss the view of well-being which serves as a foundation for my account of liberty.

What I cannot offer is a full-fledged philosophical argument for the correctness of my view of individual well-being, or for the defectiveness of rival views. With respect to the latter task, there are simply too many views; and even if I were to isolate the other major contenders and focus exclusively on them, the time that

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_5

would be required to discuss any of them fairly and thoroughly would take me far beyond the scope of my purpose in the following two chapters, which is to develop a robust, precise, and philosophically satisfying account of liberty. A theory of well-being is meant here to serve as a background against which this project may be carried out. So I shall have to be content at present with a mere sketch of the view of well-being I accept, sufficient to the purpose of providing such a background. A more thorough outline of my theory of well-being, and a discussion of the various philosophical presuppositions which lie behind it, will come in Chaps. 13 and 14—though these will still fall far short of a complete argument for the theory or an exhaustive defense of it.

The first criteria for an acceptable account of human well-being are empirical. We must exclude those which identify the worthwhile life with a manner of living that is physically or psychologically unavailable to most normally functioning adults.[1] We can narrow things down a bit further by making use of the subjective responses to survey questions which now form the basis of the burgeoning field of "happiness studies."[2] These results, however, do not get us very far. I do think it is safe to say that a mode of living which people widely report as making them miserable is no contender for the title of a worthwhile way of life, whatever theoretical considerations may be devised to support it. But the number of those living in abject poverty, or under the constant threat of domestic abuse, who claim that they are, on the whole, content with their lot, is high enough to conclude that we are fully capable of being wrong about whether or not our lives are good.[3]

As we lack the ability to determine on strictly empirical grounds what mode (or modes) of life are good and worthwhile, we must turn to the resources of ethical theory. The broad historical tradition within ethical theory I shall turn to is, unsurprisingly, the Aristotelian one. The conception of well-being that I will outline here owes much to the work of the prominent neo-Aristotelians Amartya Sen and Joseph Raz—another unremarkable fact, perhaps, given that the influence of either the one or the other can be seen in most parts of this work. The conception is the neo-Aristotelian interpretation of agent well-being given in the General Introduction:

> *Agent Well-being*: An agent's well-being consists in his willing pursuit and achievement of valuable functionings, chosen through deliberation from an adequate range of options, within the confines of respect for his moral duties.

[1] See (Flanagan 1986) for an extended discussion of the importance of psychological realism for ethical theory.

[2] For an example, see (Bruni and Porta 2007).

[3] 25 % of the rural poor in Malawi, one of the world's poorest nations, claim to be satisfied or very satisfied with their lot in life, and another 15 % claim that they are not unsatisfied (Ravallion and Lokshin 2005). 66 % of women in the Bajirah province of Ethiopia claim to believe that a man has good reason to beat his wife if she does not have the housework completed by the time he arrives home in the evening (World Health Organization 2005).

1 The Good Life and Valuable Functioning

Functionings are the activities and states of being which the agent achieves. The idea behind taking functionings as the appropriate constituents of well-being is simple—a person's well-being must be a matter of what sort of life he is managing to lead, and this means it is a matter of what he has managed, and is managing, to do and to be. The deliberative selection of which achievements to pursue is, we have seen, the exercise of the capacity of autonomy. The agent's range of options is represented by his capability set—his abilities, access to resources, and opportunities to achieve functioning. The levels of achievement, the number of functionings achieved, the extent of the capability set, the number of choices made autonomously, and the excellence of the exercise of autonomy are all potentially relevant to the goodness of the life lived. The notion of autonomy has already occupied our attention for quite some time. The notions of functioning and capability will be examined in the next chapter. I will limit myself here to some general remarks on the sense in which I take the value of functionings to be objective, and on the range of factors which can contribute to one's well-being.

Sen makes a crucial distinction between the "position relativity" of *valuations* and the "authorship invariance" of *values* (Sen 1985, p. 183). Valuations—the judgments which one reaches regarding the values of states and activities—are and should be sensitive to differences in the positions of the agents making those valuations—their differing circumstances, talents, interests, etc. So there is a sense in which valuations are subjective. Nonetheless, a valuation from a particular position is either correct or incorrect—the identity of the one making the valuation, apart from the aspects of his position which are relevant to making that valuation (which may include his tastes, interests, abilities, talents, etc.), does not matter. So the value of the state or activity to an agent in a given position is invariant, and so in a sense objective.

Sen's notion of authorship invariance dovetails nicely with Raz's anti-transparency thesis: it is not the case that nothing can contribute to an agent's well-being unless the agent recognizes it as so contributing (Raz 1986, p. 308), and it is not the case that whatever the agent believes is contributing to his well-being is so contributing (Raz 1986, p. 303). Rather, an agent's well-being depends on the actual value of his goals and achievements (Raz 1986, p. 298). Raz occasionally makes claims which appear overly subjective, such as the claim that to evaluate someone's well-being is to ask how successful his life is from his point of view (Raz 1986, p. 289). Given his commitment to anti-transparency, however, it is best to interpret "point of view" in this context as referring to the agent's position in Sen's sense.

Both Sen and Raz, and any theorist with an objective view of value, are left with the question of what sorts of functionings are truly valuable and, and how this can be determined. This is one of many points on which my work on autonomy has much to contribute. By constructing a model of excellent ends-deliberation, I have made available an answer to this question which is procedural, so to speak, rather than substantive. Those functionings, those achievements, which are genuinely

valuable from a particular position are those that could be adopted and pursued by an agent in that position as the result of a well-executed course of ends-deliberation. The assumption of authorship invariance guarantees that which functionings these are can be known by others outside the agent's position, since the answer does not depend on who the agent is beyond the aspects of himself relevant to defining his position. Given enough information about that position, the result of a well-executed course of ends-deliberation can be known to anyone.[4] So rather than providing a theory-driven list of valuable functionings, my approach provides a precise representation of the method by which we succeed in determining what sorts of states and activities are truly valuable from a given position. We should not expect to be able to identify those functionings except through careful deliberation and with adequate information about the position from which the valuation is to be made.

Both Sen and Raz self-consciously attempt to steer a middle course between subjective and objective theories of well-being, and this effort is to be both commended and imitated. Sen sees capability-based approaches to well-being as meriting the moniker "Aristotelian." For any capability-based approach focuses its attention primarily on human potential and its excellent development and exercise in valuable activity. But Sen is wary (and rightly so) of *too much* authentic Aristotelianism. Quoting Martha Nussbaum, he notes that on a plausible reading of Aristotle, Aristotle believes "that there is just one list of functionings (at least at a certain level of generality) that do in fact constitute good human living" (Nussbaum quoted in Sen 1993, p. 46). Nussbaum challenges Sen to develop a fully specified, objective, normative account of human functioning, and a complete method for evaluating the contribution of any functioning to a good human life (Nussbaum 1988, p. 176). Sen remarks that his refusal to attempt such a feat stem from his concerns that the resulting view of human life could not help but be "tremendously oversimplified," and he points out that we do not need to commit ourselves to a unique set of functionings in order to maintain that the value of functionings is objective (Sen 1993, p. 47). Behind these is concerns is, I believe, a commitment to one of the central principles of my own project, a principle which is of considerable important for political liberalism:

> *The Principle of Competitive Value Pluralism*: There are many equally good ways of life, which are incompatible insofar as leading one excludes leading others, and the values that structure some conflict with the values that structure others.

As we will see in Chap. 9 when the issue of "liberal neutrality" arises, a commitment to competitive value pluralism will prove essential to defending a conception

[4] The fact that ends-deliberation, as I have modeled, always involves *subjective* probability judgments of the choiceworthiness of potential ends might be thought to indicate that the conclusions which a given agent would reach were he to deliberate well could not be known to anyone other than that agent. But as Sen points out, subjective probability judgments are really only subjective in being position-dependent. Given enough information about the position of an agent, we can always evaluate—objectively—how reasonable the agent has been in updating his prior probabilities (Sen 2002, pp. 478–480).

of liberalism that rests on an even moderately perfectionist conception of well-being from the objections of the deontological tradition.

2 The Good Life and Liberty

A final point of agreement between Sen and Raz, and a point which I also adopt, is that there is a close relationship between a person's well-being and how free he is. Sen explicitly endorses the claim, which is obviously a pillar of liberalism, that the good life is, in large part, a life of freedom (Sen 1985, p. 202). And in discussing the range of factors which matter to an agent's well-being, Raz includes not only factors which affect the agent's pursuit and achievement of his actual goals, but also those that affect his ability and opportunity to adopt or pursue any other valuable goals (Raz 1999, p. 306).

Sen falters, however, in attempting to draw a distinction between an agent's well-being and his "agency-goals" (Sen 1985, p. 186). Sen understands the sort of functioning, or flourishing, that constitutes well-being in a relatively narrow sense, such that agent's may adopt, pursue and achieve all sorts of valuable goals that have nothing to do with their well-being, and that may even conflict with it. On Sen's view, for example, any diminishment of physical health detracts from one's well-being. Thus, if an agent is working tirelessly in pursuit of some valuable goal to which he attaches great importance, and the goal cannot be achieved without a level of effort sufficient to diminish his physical health somewhat, Sen would say that the agent is pursuing an agency-goal that is in conflict with his own well-being. Sen's attempt to distinguish between types of goals (such as good health) whose achievement always contributes to an agent's well-being, and other types of goals which, however valuable or important to the agent they may be, do not contribute to well-being (and may conflict with it), seems to me to be unmotivated. It is certainly plausible that to sacrifice one's health to some very great extent will necessarily detract from one's well-being; but this does not mean that the dedicated pursuit of any goal which requires any sacrifice of health must be placed outside the realm of well-being altogether. An agent's life is going well, as Raz says, when he is at peace with himself and is pursuing valuable goals whole-heartedly (Raz 1999, p. 310). And I would add that his life is better, at least up to a point, the greater is his freedom to choose which valuable goals to pursue.[5] But as Raz points out, there is no *essential* connection between well-being and behavior that can be fairly described as self-sacrificing (Raz 1999, pp. 315–316). We have no reason to exclude the whole-hearted pursuit of any valuable goal from the class of contributors to well-being, and no reason to deny that one can sacrifice certain aspects of one's life to such a pursuit without doing so to the point where one's overall well-being begins

[5] I do not deny that there is such a thing as too much freedom of choice, too many opportunities, and that at this point one's well-being begins to diminish.

to diminish. I therefore side with Raz in taking a broad view of well-being and of the range of pursuits and achievements which can contribute to it.

We should note, however, that the good life is a life of autonomy as much as it is a life of freedom—a life of choosing, on the basis of excellent ends-deliberation, which options to pursue from among a broad range of valuable options that constitutes a context of freedom. Since I will go on to define liberty as a compound of freedom and autonomy, I view the good life not just as a life of freedom but as a life of liberty.

3 The Good Life and the Moral Life

Finally, I concur with Raz's assertion that "One can profit from, one's well-being can be served by, compliance with, or the attempt to comply with, any moral consideration" (Raz 1999, p. 310). In my view, this is so because a normal human life is always and inevitably a life of moral agency, and I assume—I hope plausibly—that one of the necessary constituents of any good human life whatsoever is functioning well *qua* moral agent. I do not assume that an agent's well-being is diminished whenever he fails to make the best moral choice he could, all things considered. I do maintain, however, that at the very least, one's well-being is diminished when one fails to satisfy one's moral duties—this is substantial failure to function well *qua* moral agent. I will have much more to say about the nature of moral duty, and the role considerations of duty play in practical reasoning—particularly in combination with considerations about the pursuit of one's own autonomously chosen ends—in Chaps. 13 and 14.

References

Bruni, L., and P.L. Porta (eds.). 2007. *Handbook on the economics of happiness*. Northampton: Edward Elgar Publishing.

Flanagan, O. 1986. *Varieties of moral personality*. Cambridge, MA: Harvard University Press.

Nussbaum, M. 1988. Nature, function and capability: Aristotle on political distribution. *Oxford Studies in Ancient Philosophy*, *supplementary volume*, 145–154. Oxford: Oxford University Press.

Ravallion, M., and M. Lokshin. 2005. Who cares about relative deprivation? Policy Research Working Paper, World Bank, No. 3782.

Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.

Raz, J. 1999. The central conflict: Morality and self-interest. In *Engaging reason*, 303–332. Oxford: Clarendon Press.

Sen, A. 1985. Well-being, agency and freedom. *Journal of Philosophy* 82(4): 169–221.

Sen, A. 1993. Capability and well-being. In *The quality of life*, ed. M. Nussbaum and A. Sen, 30–53. Oxford: Clarendon Press.

Sen, A. 2002. *Rationality and freedom*. Cambridge, MA: Harvard University Press.

World Health Organization. 2005. *WHO multi-country study on women's health and domestic violence against women*. Geneva: WHO.

Chapter 6
The Concept of Individual Freedom

1 Introduction

My goal in this chapter and the next is to articulate a conception of individual liberty that will serve as the basis for the liberty-based account of distributive justice which I will develop in Chaps. 9, 10 and 11, and as the basis for the limitations on State intervention which I will develop in Chap. 16. I argue that we should understand liberty as a compound concept, constituted by the possession of both a developed capacity of autonomy (as this has been explicated in Chaps. 3 and 4) and a degree of negative freedom. I begin this chapter by examining the debate over rival 'concepts' of freedom—negative, positive, and so-called 'third concepts'—and argue that the positive concept of freedom is not a way of understanding freedom at all, and that the most prominent candidates for a third concept of freedom all reduce to specific conceptions of negative freedom. This leaves us with the negative concept of freedom as the only appropriate characterization of freedom. The issue then becomes one of defining a specific conception of negative freedom from which, along with my account of the capacity of autonomy, I will construct my conception of individual liberty. I conclude this chapter with a discussion of Matthew Kramer's recent "pure" view of negative freedom, and argue that while it does include some valuable insights that should be preserved, it fails on multiple counts as a complete account of negative freedom. I develop my own view of negative freedom in the next chapter.

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_6

2 Concepts of Freedom

2.1 Negative Freedom

In a very influential article, Gerald MacCallum argues that there is only one concept of freedom—the one usually denoted the 'negative' concept (MacCallum 1967). MacCallum offers the following tri-partite analysis of having a particular freedom (MacCallum 1967, p. 314):

An agent A is *free* from a constraint C to (not) do/(not) become an action/state E.

This understanding of freedom is a negative one insofar as freedom is always taken to be freedom *from* some sort of constraint or impediment. MacCallum then identifies three sources of controversy over how to understand the concept of freedom, based on the three *relata* in his analysis of freedom (MacCallum 1967, pp. 320–323). The first sort of controversy concerns the question of who counts as an agent. Agency here may be narrowly construed to include only what Anglo-American law refers to as 'natural persons,' excluding both 'artificial persons' (such as corporations) and non-human animals (which are taken to fall outside the scope of personhood). Alternatively, we may understand agency broadly, so as to include both of these groups (and perhaps others as well). The second sort of controversy concerns the kinds of things that count as constraints. On one influential view, the only genuine constraints on one's freedom are the interfering actions of other persons (and perhaps only when these actions are intentional). A less narrow view might allow for the presence of interfering factors other than the actions of others to count as constraints on freedom. And one broader still would allow not only the *presence* of interfering factors to count as constraints on freedom, but the *absence* of enabling factors as well. Finally, a third source of controversy concerns the sorts of things from which agents can be constrained when they are being deprived of a particular freedom. Again, there is an influential narrow view that claims that when one's freedom is limited, it is simply the case that there is some action that one is prevented from performing. Broader views would allow that one's freedom is limited when one is prevented from developing a trait, becoming a certain sort of person, occupying a certain state of affairs, or entering into a certain set of circumstances.

Any precise account of negative freedom must make clear what restrictions it places on the ranges of the variables in MacCallum's triadic relation. Specifying these ranges will be one important aspect of developing my own view. But for now, we may continue on to the concept of positive freedom, and determine whether there is any way to define it such that it does not collapse into some conception of negative freedom.

2.2 *Positive Freedom*

While Isaiah Berlin describes his conception of negative freedom quite clearly (it is the freedom of persons *from* the directly or indirectly interfering actions, whether intentional or not, of other persons *to* perform actions), his definition of positive freedom is not nearly so easy to lay out (Berlin 1969/2006, p. 371). He begins with the suggestion that freedom in the positive sense "consists in being one's own master" (Berlin 1969/2006, p. 373). This, as he notes, is not such a very different notion than the notion of being free from the interfering actions of others. The divergence between the negative and positive concepts of freedom is due, rather, to the rather peculiar historical development of the meaning of 'self-rule.' Let us first briefly examine how the development of this idea leads to a supposedly distinct concept of freedom. We will then be in a position to settle two issues which are important for our current purposes. First, whether it is possible to formulate a concept of freedom that falls outside MacCallum's analysis of negative freedom; and second, whether there is reason to think that a society that values and promotes self-rule on the part of its citizens is necessarily doomed to embody the sort of totalitarianism that Berlin means for his work to help us avoid.

Positive freedom, according to Berlin, is the sort of freedom that is invoked to justify coercion—the very activity which deprives one of negative freedom. The first step toward a pernicious concept of positive freedom is the observation that we can be prevented from doing what we wish, or what we think we ought to do, by the presence of urges and passions with which we do not identify, and which we cannot easily control (Berlin 1969/2006, p. 374). Such passions, then, are among the constraints which limit our available actions, and from which we can be freed. Historically, this observation developed into the idea that the self is fragmented. There is the true, rational self, and the oppressive, passion-ridden self. To be freed from the internal constraints of our passions is to defeat this lower self and liberate the rational self. The next step is what Berlin famously calls "the retreat to the inner citadel" (Berlin 1969/2006, p. 375). If freedom includes freedom from the passions which one's rational self does not endorse, then it is natural to believe that the path to freedom is the rational self's gaining control over the passions. Whenever desire serves as a constraint on reason's ability to direct the will, that constraint can be surmounted through rational control over the impinging desire, rather than through satisfaction of it. Berlin sees this idea at the core of Kantian ethics (Berlin 1969/2006, pp. 375–376). The autonomous will is not guided by inclination; it is a ruler over the inclinations that arise to confront it. This will is rational, insofar as it obeys the moral law which is discovered through pure practical reason; and it is free, insofar as it imposes this moral law on itself, and remains uncoerced by any force.

In Kantian ethics, we are left with the view that coercion is the greatest of evils, precisely because the dignity of man is due to his possessing an autonomous will,

and to coerce him is to deny this autonomy. We will have to move on from Kant, then, in order to continue the development of the positive concept of freedom. But an important Kantian idea is carried over into the next, Hegelian step toward positive freedom. This is the idea that freedom is attained through the comprehension by individual reason of rational laws (Berlin 1969/2006, p. 378). Such laws include not only things like mathematical rules, but laws of morality and social order as well. These laws set out what individual and social life must be like if individuals and societies are to function at all—that is, if they are not to operate irrationally, and thus chaotically. Reason's comprehension of rational laws is liberating insofar as it frees one from desiring the impossible—from desiring modes of individual and social life which cannot be made to function. To this idea, we must add another: that at some fairly fine-grained level of description, there can be only one rational mode of life for the individual and for society, and that these two modes of life are in perfect agreement. The rational organization of society just is the result of each and every one of its members leading a rational life (Berlin 1969/2006, p. 379).

This, then, is the positive concept of freedom. One is free in the positive sense when one is living the rational life; the mode of life that follows from comprehension of the rational xcomprehension of the rational laws of morality and society; and to live this life requires that all others in one's society also live the rational life (lest their irrational behavior disrupt the course of one's own rational life in some unavoidable way). For all the members of a society to live such a life together amounts to a rational life for that society as a whole. And from here, it is only one very short further step to the justification of coercion. Anyone who is not succeeding at living such a life threatens the very functioning of society. Though it would be best if they were able to bring themselves into conformity with the laws of reason, they may not be able to do so. So those of us who do comprehend the rational functioning of society have no choice but to bring them into conformity by force (Berlin 1969/2006, pp. 381–382). And in so doing, we are not in fact coercing them into acting in their own best interest. For we are merely compelling them to do what they would freely choose to do themselves, were they capable of liberating their true rational selves from their unruly passions. And we will succeed in making them truly free when, as a result of our coercion, these passions are finally extinguished—when we have beaten their devils out of them. We can thus claim that far from coercing them, we are rather forcing them to be free (Berlin 1969/2006, p. 382).

With a firm grip on the content of the positive concept of freedom, we can now ask whether this is really a distinct concept from the concept of negative freedom analyzed by MacCallum. Quentin Skinner does a nice job of isolating what exactly it is about the positive concept that is supposed to distinguish it so radically from the negative concept. Skinner identifies Bernard Bosanquet and T. H. Green as the two philosophers whose writings on freedom serve as the main inspiration for Berlin's explication of the positive concept (Skinner 2006). Bosanquet discusses "the 'real' or 'ideal' self whose activity is *identical* with freedom" (Skinner 2006, p. 399). Green identifies freedom as "an end-state" in which man has attained an ideal level of self-realization (Skinner 2006, p. 399). For these neo-Hegelian thinkers, the living of a fully rational life itself is the only thing that frees us from "the constraints

and obstacles to the realization of our full potential," and is thus the only thing that "makes us fully free" (Skinner 2006, p. 400). Freedom itself is then the achievement of this ideal state of self-realization that one comes to by way of leading a rational life.

Skinner concludes that Berlin has succeeded in articulating a distinct concept of freedom on the basis of the neo-Hegelian tradition:

> [T]he freedom of human agents consists in their having succeeded in realizing an ideal of themselves. But this is not to speak of a condition in which someone is free to do or become something, as required by MacCallum's analysis. It is to speak of a condition in which someone has succeeded in becoming something. Freedom is not being viewed as absence of constraint on action; it is being viewed as a pattern of action of a certain kind. (Skinner 2006, p. 400)

The neo-Hegelian conception of freedom that Berlin articulates is certainly a distinctive one, and is at quite some distance from most other influential conceptions of freedom. But Skinner is nonetheless wrong to conclude that it is really a distinct concept. It is not, as he thinks, outside of MacCallum's triadic analysis. Freedom, on the neo-Hegelian view, is a particular state which an agent can inhabit—a state from which an agent's actions will constitute leading a particular sort of life. The third variable in MacCallum's relation may range not only over actions an agent could perform or ways an agent could become; it may also range over states of the agent, ways the agent could be. The second variable, likewise, is not restricted to constraints on action. It may range over any type of constraint (to action, becoming, or being), and it may include absences as well as the presence of interfering factors. In order to analyze the neo-Hegelian conception of freedom, then, one possible path is to take the value of the third variable to be the state of ideal self-realization, and take the value of the second variable to be the trivial constraint on inhabiting this state: namely, the absence of the state itself.

But there is much that is unsatisfactory about this analysis. The main problem is the idea that we may assign trivial values to the second variable. For this would imply that there is a sense in which I am not free to perform some action whenever I am not performing it (since the doing of the action is absent), and a sense in which I am not free to become that which I am not presently engaged in becoming (since the becoming is likewise absent). The adherent of MacCallum's analysis has two options here. First, he could accept that the analysis allows for trivial conceptions of freedom, according to which one is free to perform some action (or what have you) only so long as the doing of that action is not absent from the agent—that is to say, only so long as the agent is performing that action. We would then, of course, be entitled to say that this conception of freedom is not a useful or significant one, is not what anyone actually means when they speak of freedom, etc. Some might argue, however, that an analysis of the concept of freedom that allows for such conceptions must be defective in some way. If this is right, then the range of the second variable must be limited so that it cannot take as a value the mere absence of whatever occupies the third place in the relation. This is the second option. If we accept this restriction, and so rule out the trivial conceptions, can we still accom-

modate the neo-Hegelian notion of freedom within the framework of negative freedom? I believe that we can.

If the absence of the state of ideal self-realization is always due to some other interfering factors, then these interfering factors can be identified as the constraints which prevent self-realization (rather than identifying the mere absence of that state as the relevant constraint). And the neo-Hegelians do seem to think that there are distinct constraints that prevent us from attaining self-realization: namely, the irrational passions. So we might be tempted to conclude that for the neo-Hegelians, when we are free from the last vestiges of irrational passion, that is when we are free to attain self-realization, and the freedom to attain self-realization is the conception of individual freedom that they subscribe to. In fact, the issue is somewhat more complicated than this. Once one has relinquished all irrational passion, and is thereby actually succeeding in living a life governed by reason and the comprehension of rational laws—a life of contributing to the collective human project of working out the content of, and realizing, the common good—one is by that very fact leading a life of self-realization. That is just what a life of self-realization—a life in accordance with our essentially social human nature—is, on this account. So it is not as if self-realization is something that would finally be open to us once our reason had conquered all of our irrational passions (waiting only for us to reach out and grab it); rather, it is something we would have attained in virtue of that victory. As Skinner explains, "if and only if we *actually follow* the most fulfilling way of life [that is, a fully rational one] shall we overcome the constraints and obstacles to our realization of our full potential, and thereby realize our ideal of ourselves. *The living of such a life alone* frees us from such constraints [emphasis added]" (Skinner 2006, p. 400).

What this suggests is that for the neo-Hegelians, freedom from the constraints of irrational passions is freedom not to *become* self-realized, but rather to *be* self-realized. Once we are free from passion, we are free to inhabit the state of ideal self-realization. To this conception of freedom, the neo-Hegelians add the claim that there is no distinction to be drawn between being free to inhabit this state and actually inhabiting it—no distinction between being free to *be* self-realized and *being* self-realized. But this just falls out of the way they understand self-realization. If to be self-realized is just to win the final victory of reason over passion, then it follows that the state of self-realization will be achieved at the same moment that the last trace of the constraints on self-realization vanish—one becomes self-realized precisely when one becomes free to be self-realized. The neo-Hegelians, in other words, have defined the third relatum in the freedom relation in such a way that its achievement follows directly from the absence of the second. And it is this fact that makes intelligible their definition of freedom itself as an end-state, as the actual achievement of self-realization. Their conception of freedom is the freedom from irrational passion to be self-realized, and there is no space, given their account of self-realization, in which to drive a wedge between being self-realized and being free to be self-realized: "Liberty *consists in* following that way of life in which, all passion spent, we finally achieve harmony with nature [emphasis added]"—i.e. our human nature as social animals (Skinner 2006, p. 400).

I conclude, then, that even the neo-Hegelian notion of freedom can be brought within the concept of negative freedom as analyzed by MacCallum: An agent A is free when he is free from the constraints of irrational passion to be self-realized. The nature of the state of self-realization makes it the case that one is free to be self-realized if and only if one is self-realized. The so-called positive concept of freedom has proved to be one conception of negative freedom among many. Let us turn, then to the second question Berlin's work has left us: are we doomed to walk the path toward totalitarianism if we embrace a conception of freedom that diverges from his own and shares the concerns of the tradition that produced the neo-Hegelian conception?

Recall that the positive tradition began by emphasizing the importance of the freedom to exercise self-rule. One can hardly argue that there is anything pernicious about this notion of self-rule in itself. It is a plain fact of life that we all find within ourselves wants, urges and temptations for things which we firmly believe are not good for us. Likewise, we all find ourselves subject to pressures and influences from our societies and cultures which push us in directions we do not judge choiceworthy. And we are better off for not simply choosing to follow these urges and pressures wherever they lead. To determine for oneself what goals are worth pursuing, and act on that determination, is an important aspect of a valuable life. To accept this does not commit one to the absurd claim that one's freedom increases every time one successfully eliminates a preference for something one cannot presently attain. For the preference may be a deliberate one, rather than one resulting from an unwelcome urge or influence.[1]

So where exactly did the positive tradition go wrong? In tracing the history of this tradition, Berlin rightly identified Kant as the pivotal figure. For Kant, a free person is a person possessing an autonomous will: a person who chooses his pursuits *via* the exercise of his reason, rather than simply following his inclinations (or, we might add, his socio-cultural influences). This understanding of freedom was supposed to make the evil of coercion perfectly evident: coercion is just that activity which violates an individual's autonomy, which is the basis of his dignity and of the respect that is owed to him. So far, so good. But from this point, a path to justifying coercion in the name of freedom opens up through the Kantian understanding of reason. It is by understanding the exercise of reason in terms of the discovery of moral and social principles of universal validity that we start down this path.

The pernicious turn in the positive tradition, then, comes at the moment when the rational life is understood in terms of a set of universally valid moral and social principles which is sufficiently comprehensive to exclude the possibility that a society may embrace competitive value pluralism and still function well. Recall that one of the liberal principles to which we are committed is:

> *The Principle of Competitive Value Pluralism*: There are many equally good ways of life, which are incompatible insofar as leading one excludes leading others, and the values that structure some conflict with the values that structure others.

[1] John Christman makes a similar point (Christman 1991, p. 353).

According to competitive value pluralism, the multitude of good forms of life are not all perfectly compossible. It may be that the achievement of a particular goal by one person leading one form of worthwhile life is incompatible with the achievement of another goal by another person leading a different but equally worthwhile form of life. The positive tradition post-Kant rules out this possibility, by asserting that the universal moral and social principles that structure a rational life are extensive enough to allow for only one form of rational life for all individuals, and only one form of rational life for any society (which is realized when all the members of a society lead rational lives).

So it is not in valuing self-rule that the positive tradition courts totalitarianism. Rather, it is in rejecting competitive value pluralism as a result of adopting a particular understanding of the exercise of practical reason. Our question then becomes: can we value self-rule without committing ourselves to this post-Kantian view of practical reason? It seems clear that we can. Self-rule is just the ability and opportunity to act for one's own purposes and on one's own reasons. The characterization of autonomy I have developed in Part I is just as compatible with this notion as the Kantian characterization of autonomy is. But on my view, the process of deliberating rationally about ends need not take the form of discovering any universally valid moral and social principles. The agent deliberates about what he should judge choiceworthy based on his own situation, interests, talents, etc. His deliberation proceeds by incorporating evidence from a broad range of sources. There are many equally reasonable conclusions he may come to, and his conclusions may differ significantly from the conclusions of another, differently situated agent without being any less reasonable. So my account of autonomy as the capacity for ends-deliberation is perfectly consistent with competitive value pluralism.

There is reason to think, in fact, that the development of autonomy as I have characterized it requires a pluralistic society. In evaluating the choiceworthiness of a potential goal, one of the sources of evidence the autonomous individual relies on is how well pursuit of that goal would cohere with his other values, commitments and pursuits. Stanley Benn characterizes the autonomous person as one who is committed to a "critical, creative, and conscious search for coherence within his system of beliefs," where we may take an agent's system of beliefs to include his preferences (as we understand a preference to be a belief about the choiceworthiness of an option) (Benn 1988, p. 179). Benn argues that only within a pluralistic society can an individual pursue coherence in the way required for autonomy. The traditions of a society or culture may present its members with an extensive and coherent ready-made system of beliefs, but those who adopt the traditional views will not be autonomous in virtue of being coherent. Autonomy requires that one arrive at one's commitments deliberately. This means weighing the considerations that favor different commitments, which in turn requires that the agent be sensitive to those contrasting considerations. That sensitivity needs a pluralistic environment in order to develop and flourish; otherwise, individuals are unlikely to receive the necessary exposure to differing sets of commitments and the considerations that favor them. "A monolithic system…would simply lack the incoherences which leave space for

autonomous development. Where there is no work to be done, none can claim credit for doing it" (Benn 1988, p. 182).

My theory of autonomy, then, provides a way of understanding the idea of self-rule which makes the value of this attribute clear. The theory does not justify the "retreat to the inner citadel" according to which even deliberate and reasonable preferences ought to be jettisoned if they are difficult to satisfy. It is consistent with competitive value pluralism and even requires a pluralistic society in order for individuals to develop the capacity of autonomy. It thus avoids the mistake which originally led the positive tradition down the path to totalitarianism. It should be clear that as I have characterized autonomy, the very idea of coercing someone into being more autonomous is nonsense. If one has been coerced into adopting a preference, then one has not adopted that preference on the basis of one's own deliberations.[2] Is my theory of autonomy, then, itself a theory of freedom in the positive tradition? It is not. Unlike many in the positive tradition since Kant, I do not take my theory of autonomy itself to articulate the conception of freedom that is fit to play the role of a guiding ethical and political value. I do not, for that matter, take it to articulate any conception of freedom at all. Freedom and autonomy are separate concepts. But they are closely related, and as we shall see, my theory of autonomy will guide my development of a conception of freedom. But before I proceed to argue that freedom and autonomy should be kept conceptually distinct, and to discuss the important relationship between these two concepts, I will examine some recently proposed "third" concepts of freedom. If they too prove to be additional conceptions of freedom falling within the negative concept, we should be content, at least for the time being, that one concept of freedom is sufficient.

2.3 "Third" Concepts of Freedom

Samuel Fleischacker takes negative and positive freedom to be two distinct concepts of freedom. He draws the distinction between them along Kantian lines. Negative freedom is the freedom to do what one desires to do, or to follow one's inclinations (Fleischacker 1999, p. 253). Positive freedom is the freedom to follow the dictates of one's rational will (Fleischacker 1999, p. 253). He sees both types of freedom as defective in various ways, and seeks to articulate a third, by appealing to the third Kantian faculty: judgment. He describes judgment as "a 'mediating' faculty, by which I may identify myself either with my desires or with my will" (Fleischacker 1999, pp. 253–254). He identifies the exercise of judgment with autonomy, or self-rule: "For rule over oneself is quintessentially the exercise of judgment or phronesis, the making of choices guided by judgment" (Fleischacker 1999, p. 251). As in Benn's characterization of autonomy, the gradual search for

[2] Nor can an agent be coerced into exercising his capacity to deliberate about his preferences, though this is more of a practical matter than a conceptual one. The presence of any significant coercion would doubtless interfere with the process of rational deliberation of ends.

coherence among one's commitments is an important part of exercising judgment (Fleischacker 1999, p. 257). Fleischacker's third concept of freedom, then, is the freedom to develop, exercise, and act on one's capacity for judgment. One crucial aspect of this type of freedom is the freedom to "change our desires in a reasonable way, and not merely to act on whatever desires we already happen to have" (Fleischacker 1999, p. 254). In so changing our desires, however, the bounds of what is reasonable are not drawn by the dictates of the rational will. The exercise of judgment is also "open to the influence of cultures or traditions" (Fleischacker 1999, p. 255).

Like many philosophers in the positive tradition since Kant, Fleischacker identifies the ethically and politically relevant conception of freedom as the freedom to develop and exercise the capacity of autonomy, and to act autonomously. He identifies the capacity of autonomy as the capacity for judgment in a Kantian sense, but there is good reason to think that what he has in mind is something close to my characterization of autonomy as the capacity to deliberate rationally about ends. His emphasis on the reasonable revision of preferences, and the relevance of a broad range of evidence including cultural traditions to this process, bring out the similarity nicely. His view, however, suffers from two important defects, one conceptual and one of content. First, there is no reason to delimit distinct concepts of freedom along such narrow Kantian lines. Fleischacker's purported third concept of freedom is actually one more conception of negative freedom, the only concept of freedom we require. It is the freedom from certain constraints (at least including the interfering actions of other people) to exercise, develop and act on a particular capacity. Second, Fleischacker's characterization of the faculty of judgment is lacking in content. As we shall see below, my own conception of negative freedom is fairly close to Fleischacker's. But he has failed to develop a precise theory of how an agent would go about making reasonable revisions to his preferences, taking into account social and cultural factors, emotional responses (which Fleischacker might classify as desires or inclinations) and other sorts of relevant considerations along with categorical reasons. Constructing such a theory is precisely what I have done in Part A. We may move on from Fleischacker, then, recognizing that he has managed to focus our attention in the right area as we search for an adequate conception of negative freedom, and that he has seen the importance of a theory of autonomy of the kind I have developed.

The other potential candidate for a third concept of freedom is the notion of "republican freedom" articulated in the work of Quentin Skinner and Philip Pettit (Pettit 2001; Skinner 2006). Skinner identifies this concept of freedom as the one adopted by the Republican opponents of the King's unlimited prerogative in late seventeenth century Britain (Skinner 2006, p. 403). The basis for the unlimited royal prerogative is the idea that the King is the fount of all authority. He then has the power to apportion subsidiary powers as well as duties, rights, immunities and liberties as he sees fit. For an Englishman to have a liberty, then, is simply for it not to be the case that the King has placed a particular duty on him. But if the royal prerogative is unlimited, the King may cancel the liberties of his subjects at any time. His subjects would have no immunities against the exercise of his power. The

Republican opposition, which believed in certain inherent rights and liberties of Englishmen, disputed the idea that the royal prerogative is unlimited. They claimed that the people did not really possess the liberties they were entitled to, so long as the King had the power to revoke those liberties. It is this second-order freedom, this independence from authority for the preservation of one's first-order freedoms, that Skinner offers as a distinct third concept of freedom.

We can, however, perfectly well understand this notion of independence from authority as an instance of negative freedom. Recall that MacCallum's analysis allows for absences to count as constraints. Those subject to an unlimited royal prerogative lack legal immunities against the King's exercise of power. In lacking such immunities, they are placed under a constraint which deprives them of the freedom to do all manner of things. Persons in such a position are not really free to do or say anything which will lead the King to revoke their purported freedom and then punish them. This was the substance of the Republicans' complaint. It was meaningless to say that they possessed certain freedoms, if exercising those freedoms in ways that displeased the Crown would lead to the revocation of those purported freedoms and to punishment for the offending acts. In the terms of MacCallum's triadic analysis, the English people lacked the freedom *from* the absence of legal immunities against the Crown *to* exercise their other (purported) freedoms in ways that might displease the Crown.

Republican freedom, then, is not a distinct concept of freedom. It is, however, an extremely important instance of negative freedom from a moral and political perspective. If one does not possess republican freedom—that is, if one does not possess an immunity against the arbitrary exercise of the power to limit ones other freedoms—one can hardly be said to have those freedoms in the first place. Republican freedom then makes possible the possession of other freedoms in a way that other instances of negative freedom do not. The lesson to be drawn then is that any adequate conception of freedom as a guiding social and political value must include republican freedom. It is important to note that republican freedom does not require that one's other freedoms be unlimited or absolute. It requires only that one have an immunity against the arbitrary exercise of power to limit one's freedom. Limitations on freedom through due process are consistent with republican freedom, so long as one retains an immunity against infringement on those freedoms which have not been so limited.

3 Christman on Autonomy as Positive Freedom

We have examined a number of proposals for how to articulate a concept of freedom distinct from the negative one, and have found that none succeed. We may conclude then, that at least until some new proposal is made, a single negative concept of liberty will suffice. But in examining these other proposals, we have learned some important lessons. The task that lies ahead is that of articulating a conception of negative freedom which is fit to play the role of a guiding ethical and political value.

We have found two instances of negative freedom which must be part of any such conception. First, freedom from impediments (of a still-to-be-specified range) to develop, exercise, and act on the capacity of autonomy. And second, freedom from susceptibility to arbitrary authority to exercise one's other freedoms. Before we proceed to the task of articulating a conception of negative freedom that incorporates these elements (and others), there is a question left over from the discussion of the positive tradition that must be answered: should we regard the concept of autonomy itself as a distinct positive concept of freedom? This question provides an opportunity to discuss the relationship between the concepts of freedom and autonomy, and the ways in which freedom and autonomy are themselves parts of a larger ethical and political value.

John Christman identifies positive freedom with autonomy, or more precisely, with autonomous preference-formation and preference-change (Christman 1991). Christman articulates a set of conditions which are meant to articulate autonomous preference-formation and change (Christman 1991, p. 347), and takes these conditions to be constitutive of positive freedom (Christman 1991, p. 359). Christman requires (a) that the autonomous agent be in a position to reflect on the process that generates his preference; (b) that he not resist that process (or would not were he to attend to it); (c) that his non-resistance is not the result of any factor that inhibits self-reflection; and (d) that the process of self-reflection by which he decides whether to resist the process of preference-formation be self-consistent, and the preference that results be minimally rational—e.g. that the newly generated preference fit into the agent's preference-ranking in a way that preserves transitivity. Christman, like Fleischacker and myself, sees autonomy as fundamentally a matter of how an individual comes to have the preferences he has (Christman 1991, p. 346). And like Fleischacker, he fails to provide a rigorous treatment of how this process of reflecting on and modifying one's preferences might proceed—though he does affirm that the autonomous agent's process of self-reflection is a deliberative one (Christman 1991, p. 348).

Christman articulates his conditions in a way that allows the process of preference-formation itself to be strictly causal and non-deliberative. His agent may be understood as observing a process in himself which he did not initiate, and deciding whether or not to intervene. He very quickly, however, makes it clear that in general, preferences formed in this strictly causal way cannot be autonomous, since the agent would likely not be able to resist them even given a detailed knowledge of the causal process that was occurring (Christman 1991, p. 348). Christman's difficulty on this point stems from his failure to separate the concept of a preference—a judgment about what is choiceworthy—from the philosophical concept of a desire—which combines the concept of preference with that of non-cognitive affect-response.[3] If he had kept preference and affect separate, he would not have phrased his conditions for autonomy so as to allow preferences to simply befall agents, and then had to retreat from the idea that preferences formed in this way without any place for deliberation could count as autonomous. Autonomous preference-formation,

[3] See Chaps. 3, notes 5, 6, and 4, note 14.

as we saw in Chap. 4, is a synthesis of affective attachment and deliberative judgment about the value of the things to which one is attached; and autonomous first-order preferences are the judgments about how to rank potential ends which one arrives at through this synthesis.

But Christman's understanding of autonomy is close enough to mine that we must examine his argument for identifying autonomous preference-formation and change with a distinct concept of positive freedom, and determine whether it succeeds. Christman's argument relies on the familiar example of the "oppressed housewife" (Christman 1991, pp. 344–345). Imagine a woman who grew up in a culture which strongly inculcated the idea that women should not pursue enterprises or activities outside their homes. Such a woman is clearly not autonomous—her preferences have in a sense been forced on her, and she lacks the ability to modify them on the basis of her own deliberations. Suppose she is then transplanted into a society in which women enjoy all the same opportunities as men. Christman claims that this woman would then enjoy a great deal of negative freedom, but would nonetheless remain unfree in an important sense. She would be severely constrained by her own preferences, in whose development she had no say.

This example, however, is not sufficient grounds for identifying autonomy with a distinct positive concept of freedom. Let us examine this situation more carefully. Initially, this woman's negative freedom is clearly limited—there are numerous valuable opportunities which her culture denies her. But one of the most important instances of negative freedom which she lacks is the freedom to develop the capacity of autonomy. The pressures and influences of her culture act as constraints on this process of development which make it practically impossible. As a result, she reaches adulthood without having developed the capacity of autonomy, and so is not an autonomous individual. This is what explains her inability to act autonomously once she enters a freer society—she continues to lack the capacity of autonomy. This should not surprise us; we would not expect the change in cultural climate to magically invest her with this capacity. Does she lack freedom in any important sense in her new environment? If we imagine her new society as one which merely affords the freedom to act autonomously to those who happen to have the ability to do so (as Christman seems to do), then the answer is yes. But in that case, the freedom she lacks is the freedom to develop the capacity of autonomy. If her society is one which provides her with the resources she needs to develop this capacity, and helps her to acquire the ability to transform those resources into a developed capacity, then she does have the freedom to become autonomous. This sort of freedom—freedom from the lack of the resources and abilities needed to achieve some way of being—is often called "effective freedom" or "capability-freedom." This is an important aspect of negative freedom, which I discuss in the next chapter. We have already seen that, should she become autonomous, she will have the freedom to exercise and act on her autonomy. So assuming that she now has the effective freedom to become autonomous and to exercise and act on her capacity of autonomy once it is developed, in what important sense does she remain unfree?

It remains possible that, despite these effective freedoms, she will not in the end develop the capacity of autonomy, or will not exercise it if she does, and will

continue to live as she did in her former oppressive culture. We could then interpret Christman as claiming that in a case like this, she remains unfree insofar as she is not living autonomously, despite having all the negative freedom she could hope for. This would give Christman's position the same structure as the neo-Hegelian position. There would be an end-state in which one had developed and was exercising the capacity of autonomy, and positive freedom would be identical with attaining this end-state. So long as one was not actually autonomous, one would not be positively free. The neo-Hegelian state of self-realization, however, was a state which was blocked by identifiable constraints: the irrational passions. The reason it made sense to identify being in this state as one way of being free was that one entered this state as soon as one became free to enter it—as soon as the irrational passions which constrained one's freedom to enter this state were overcome. As we saw, there was no room to drive a wedge between being free to be self-realized and being self-realized. But the state of being autonomous—whether we take that to mean having the capacity of autonomy, exercising it, or acting on the conclusions of its exercise—it not like this. One can be free to develop or exercise autonomy without doing so. And if the woman in our example simply fails to occupy the state of being autonomous, then this does not imply the presence of any constraint that prevents her doing so—unlike in the neo-Hegelian case, in which the failure to be self-realized automatically implies the presence of irrational passions. So it does not make sense to speak of the mere fact that she has not actually developed the capacity of autonomy as a way of being unfree. If there were some constraint on her becoming autonomous, then she would lack the effective freedom to be autonomous; and again, I shall argue in the next chapter that effective freedom is a crucial aspect of negative freedom, just as I argued that freedom from irrational passion is an instance of negative freedom.

It would be unjustified for Christman to simply insist, then, that failing to be autonomous is one way of being unfree. The concepts of freedom and autonomy are distinct, and should be kept so. Nonetheless, Christman would be right to point out that the life of the woman in our example is far from the ideal of the liberal society. And she falls short of this ideal precisely insofar as she fails to live autonomously despite being free, in every relevant sense, to do so. So the lesson to be drawn from Christman's discussion is, I think, not that autonomy constitutes a distinct concept of freedom, but rather that freedom—in the negative, and only, sense of the word—cannot by itself be the central value of liberalism. Autonomy also plays a central role. But freedom and autonomy are intimately connected in the liberal ideal of the good life. This is the ideal of a life spent pursuing and achieving valuable ends which have been autonomously chosen from a broad range of options which one has the freedom to pursue. This close connection between autonomy and freedom in the liberal ideal of the good life is the reason I suggest that we distinguish between freedom and liberty, and reserve the term liberty to refer to this joint realization of freedom and autonomy. In Chap. 11, when I argue that the distribution of liberty in this sense is the proper distributive concern of the State, we will see that in addition to securing every individual's freedom to develop and exercise autonomy, the State should have as one of its goals the active encouragement and incentivization of

autonomy development. A just State, then, will turn out to be one that takes measures beyond securing individual freedom to minimize the chances that the kind of situation Christman considers will arise.

4 Negative Freedom: Against Kramer's "Neutral" View

Matthew Kramer has recently developed a thorough conception of negative freedom (Kramer 2003). I will briefly discuss Kramer's view before proceeding to develop my own conception. Kramer's view suffers from two sorts of difficulty: it has a number of internal problems, and it is fails to articulate a conception of freedom that could serve as a guiding moral and political value.

Kramer states that his goal is to provide a politically and morally neutral analysis of the concept of individual freedom (Kramer 2003, p. 4). This goal must be interpreted carefully. Kramer does see freedom as an appropriate object of moral and political concern. He acknowledges that we ought to care about how much freedom people have, and that we should sometimes act for the sake of increasing the freedom of others. But he wishes to carry out an exercise in conceptual analysis that does not rely on any of the normative commitments of any particular moral or political theory. To this end, he focuses "not on freedom as a normative condition, but on freedom as a physical fact" (Kramer 2003, p. 4). He thus offers the following definitions of being free, unfree, and not-free:

> A is free to ϕ iff it is physically possible for A to ϕ.
>
> A is unfree to ϕ iff (1) A is directly or indirectly prevented from ϕ-ing by some action(s) or disposition(s) to act of some other person(s) and (2) A would be able to ϕ in the absence of (1).
>
> A is not-free to ϕ iff it is not physically possible for A to ϕ.(Kramer 2003, p. 5)

Kramer extends the definition of 'physically possible' to include what is mentally and psychologically possible, as mental and psychological states supervene on physical ones (Kramer 2003, p. 265). So an agent will not be free to do what he cannot do on account of a psychological barrier or a mental limitation (including being in a state of ignorance), even if his body is capable of performing the required movements.

The following five points should help to fill-out and clarify this basic account. First, Kramer counts an agent as free to eschew X so long as it is physically possible for the agent to do so, even if the agent is "irresistibly obliged" to do X (Kramer 2003, p. 17). Second, there is no place for the concept of preference in the account. An agent's freedom is solely a matter of what he is physically capable of doing, regardless of his judgments of the choiceworthiness of his actions (Kramer 2003, p. 33). Third, Kramer allows that one may be free or unfree not only to perform actions, but to become/be/remain/undergo states and events (Kramer 2003, p. 158). Fourth, Kramer denies that particular freedoms can come in degrees. Whether or not it is physically possible for one to ϕ is not a matter that admits of degrees

(Kramer 2003, p. 169). Fifth, Kramer offers the following ratio as a measure of the extent of an individual's overall freedom: $F^2/(F+U)$, where F is equal to the number of freedoms the individual has, and U is equal to the number of unfreedoms (i.e. the number of freedoms the individual lacks on account of the actions or dispositions of others) (Kramer 2003, p. 359).

Having set out Kramer's view, we may proceed to a brief discussion of its most significant flaws. The first concerns the very idea that freedom can be exhaustively understood in terms of what it is physically possible for an agent to do, and so strikes right at the heart of the account. Suppose I am walking through Benjamin Bannekar park in Washington, D.C., and I happen to spot General David Petraeus walking his dog. Let us pose the following question: am I free to walk up to General Petraeus and order him to withdraw all U.S. military personnel from Iraq? I am physically capable of walking up to the General, and having done so, of uttering the words "I order you to withdraw all military personnel from Iraq" loudly enough for him to hear them. But this does not amount to giving the order—only to aping the speech-act which would constitute giving the order were that act to be performed by the President, rather than myself. I cannot order the General to do anything, because I lack the requisite authority.[4] And so it seems plausible to say that I am not free to give the General this order; that act is not among the things that I can do. But lacking the requisite authority is not a physical (or mental or psychological) inability of any kind. It is a matter of certain institutional facts failing to obtain. If, through some extraordinary chain of events, the President were to invest me with the authority to issue orders to General Petraeus, this act of the President's would not amount to making it physically possible for me to do anything that I was not physically capable of doing before. The moral here is that a physical description of an agent's movements, no matter how detailed and precise, is often insufficient to answer the question of whether or not he has performed an action; and a description of his physical abilities is often insufficient to answer the question of whether it would be possible for him to perform an action. Answering these questions often requires knowledge of the background of institutional facts against which the agent's motions take place. Kramer's account is thus blind to one of the most important aspects of human action in social environments.

The second major internal problem with Kramer's account is that he fails to justify his analysis of unfreedom. He refuses to allow that the omissions of others can be sources of an agent's unfreedom, on the grounds that doing so would make it impossible for the agent's own actions to result in unfreedom (Kramer 2003, p. 342). For if an omission were a source of unfreedom in a case where an agent's own action leaves him unable to do something, then performing the omitted action would have prevented the agent from rendering himself unable. There is a ready response to this concern, and Kramer is aware of it. We could allow omissions of actions as sources of unfreedom only if those actions should have been performed, all things considered (or, more restrictive still, only if those actions were morally

[4] In Chap. 15, I address the question of what makes it the case that an individual possesses legitimate practical authority.

obligatory). Kramer's reason for rejecting this option is totally unsatisfactory. He rejects it solely on the grounds that it would require a moral theory to determine whether an omitted action should have been performed, or was morally obligatory. It would thus spoil his efforts to provide a morally neutral analysis of the concept of freedom. But the fact that recognizing certain omissions as sources of unfreedom would spoil his project is hardly any reason to conclude that we would be wrong to recognize them as such. Rather, it is a reason to think that we have reached a point at which the integrity of Kramer's project begins to break down. And given how problematic is his understanding of freedom as purely a matter of physical fact, it is far from clear that we should make any extraordinary sacrifices in order to save the project.

I take Kramer's rejection of a preference-based account of freedom to be a third problem with his account, as it rests on the erroneous view that freedom is of independent value. But I shall not discuss this point here. Kramer adopts this position from Ian Carter, and I will discuss the concept of independent value and Carter's view of the value of freedom in Chap. 16.

A fourth problem with Kramer's view concerns his proposal for measuring the extent of an individual's freedom. Recall that Kramer offers the following ratio as a measure of the extent of an individual's overall freedom: $F^2/(F+U)$, where F is equal to the number of freedoms the individual has, and U is equal to the number of unfreedoms (i.e. the number of freedoms the individual lacks on account of the actions or dispositions of others) (Kramer 2003, p. 359). The idea that the extent of one's freedom should be measured by a ratio, which Kramer adopts from Hillel Steiner, has some plausibility. The ratio $F/(F+U)$ would express the extent of an individual's freedom by comparing the number of things he is free to do with the number of things he would be free to do but for the actions of others (Kramer 2003, p. 358). Kramer recognizes that this ratio is problematic, insofar as it would attribute equal freedom to an agent who was free to do one thousand things and unfree to do another thousand, and an agent who was free to do one thing and unfree to do one other. Kramer's way out of this problem is to square the numerator. But as Martin van Hees and Keith Dowding have pointed out, this is an utterly *ad hoc* solution (van Hees and Dowding 2009). Kramer claims that by squaring the numerator, we recognize that in judging the extent of someone's freedom, we ought to place a greater emphasis on what the person is free to do than on what he is unfree to do (Kramer 2003, pp. 368–369). First, it is unclear why we ought to do this. Second, even if this were something we ought to do, there is no reason to think that the right way to do this is by squaring the number of freedoms in the numerator of the ratio. Why not cube it, or raise it to the power of 3/2? Kramer cannot answer these questions, because he does not derive his measure in any principled way. He simply takes Steiner's measure, which is intuitively appealing, and devises a quick fix for one of its most significant shortcomings.

Let us proceed, finally, to a difficulty of another sort which besets Kramer's account. The conception of freedom he develops is unfit to play the role of a guiding moral and political value. I do not take this to be an internal flaw of his view, despite the fact that he intends his analysis to preserve an important place for the concept of

freedom in moral and political discussion. If the account did not suffer from the other problems just discussed, it would be a perfectly fine conception of freedom, despite its inability to play this role. It would simply fail to be the conception we are looking for, if we are looking for a conception that can play this role. But since Kramer believes it is fit to play this role, it is worth briefly discussing why it fails on this front.

There is some sense in making a distinction between those things that an agent is unfree to do and those that he is not-free to do. There are things that no human being is capable of doing, and, probably, things which no sane person would want to do. Being unable to do such things is not something that should concern us. So if we refer only to such things as ones we are not-free to do, then in judging the extent of an individual's freedom (which I take to be a morally and politically important feature of an individual), we are right to concern ourselves only with the individual's freedoms and unfreedoms. But as we have seen, Kramer does not define unfreedoms in such a way that only actions of these types are excluded. He defines unfreedoms as freedoms an individual lacks on account of the actions or dispositions of others. Even if he were willing to include freedoms an individual lacks on account of omitted actions whose performance was morally obligatory (or even just morally right), his account of unfreedom would still be severely flawed. And this flaw is what makes his account ideologically defective in addition to being internally problematic.

If we are interested, from a moral and political point of view, in how much freedom individuals have (and both Kramer and I think we are), then we are interested in the distribution of freedom. And to be concerned for the distribution of freedom, like a concern for the distribution of anything else (be it resources, wealth, happiness, rights, opportunities or what have you), is to be concerned with justice. The world is filled with natural injustices—disadvantages which befall individuals through no doing of their own or anyone else's. There is simply no principled reason why we should treat these natural injustices any less seriously than other disadvantages—there is nothing inherently just about the natural lottery whatsoever. So the fact that an individual lacks a freedom on account of bad luck, rather than on account of the action or blameworthy omission of himself or another person, is no reason, from the perspective of one interested in justice, to discount that lack of freedom in judging how much freedom that individual has.

But this is precisely what we do in following Kramer's account. One may lack freedoms of great importance to one's ability to lead a worthwhile life, without that lack having any direct impact on the extent of one's freedom. There will be an indirect impact, insofar as the total number of one's freedoms will be smaller than it would otherwise be. But this indirect impact is minimized, given the fact that Kramer squares the numerator of his ratio, and the fact that a smaller number of total freedoms makes for a smaller value in the denominator. Kramer's understanding of freedom and unfreedom, therefore, fails to capture both the importance of freedom (and the lack thereof), and the nature of our concern for individual freedom from the perspective of an interest in justice. His account thus fails to articulate a

conception of freedom fit to play the role of a guiding moral and political value. Success at this task is the aim of the next chapter.

5 Conclusion: The Conservative Conception of Freedom

In the Historical Introduction, I identified the conservative conception of freedom as the freedom to dispose of one's property as one wishes. Before moving on to construct my own conception of freedom, I must say something about how I plan to evaluate this proposal. The first point to be noted is that this is only a plausible candidate for a valuable form of freedom if it is restricted to one's justly held property, and to ways of disposing of one's property which do not violate the moral duties one owes to others. In the context of modern conservatism, this means the freedom to dispose of property acquired without force or fraud through activity in a free market. Arguments for endorsing the conservative conception of freedom as our guiding moral and political value are rarely direct attempts to establish that this is the most valuable form of freedom a person can have—and at any rate, the next chapter's construction of a genuinely and universally valuable conception of freedom will make any such attempt moot. Rather, the conservative conception of freedom is normally defended *via* arguments, made on other grounds, in favor of a free market maintained by a minimal State. There are three such arguments. The first claims that unique economic benefits are realized by a purely free market system; the second claims that it is under such a system that every individual receives his deserved share of the rewards of activity; and the third claims that any other system of social organization entails the violation of individual moral rights. At the end of Chap. 10, we will be in a position to see that none of these arguments succeeds. In particular, we will see that there is no account of what it means to hold one's property justly which is consistent with defending the free market and the minimal State, and thus the conservative conception of freedom.

Rather, one's property is justly held if, and only if, it is acquired in the context of participating in a society of equality of liberty. The relevant conception of liberty will be defined at the end of the next chapter, and the theory which identifies equality of liberty, in that sense, as its goal, will be defended in Chap. 11. Chapters 13, 14, 15 and 16 will argue that each member of a society has a moral duty to contribute to the creation and preservation of the conditions of equal liberty, and that the State may justifiably enforce this duty. One's justly held property, then, is the property one acquires through participating in a society of equal liberty, *which participation includes* making the contributions to maintaining such a society which one is duty-bound to make. The liberal need not and should not deny the value of being free, inside the limits of one's duty, to dispose of one's justly held property as one wishes. But recognizing this means no more than recognizing that the limits on the State's justifiable authority to restrict the actions of individuals extend no further than what is required to achieve the goal of creating and maintaining a society of equal liberty—as I will argue at the end of Chap. 16. The remainder of the book may

thus be taken as an extended argument that the freedom to dispose of one's property as one wishes is entirely subordinate to the goal of achieving an equal distribution of individual liberty, as this notion is defined in the next chapter.

References

Benn, S. 1988. *A theory of freedom*. Cambridge: Cambridge University Press.
Berlin, I. 1969/2006. Two concepts of liberty. In *Contemporary political philosophy: An anthology*, ed. R.E. Goodin, and P. Pettit, 369–386. New York: Blackwell.
Christman, J. 1991. Liberalism and individual positive freedom. *Ethics* 101(3): 343–359.
Fleischacker, S. 1999. *A third concept of liberty*. Princeton: Princeton University Press.
Kramer, M. 2003. *The quality of freedom*. Oxford: Oxford University Press.
MacCallum, G. 1967. Negative and positive freedom. *Philosophical Review* 76(2): 312–334.
Pettit, P. 2001. *A theory of freedom: From the psychology to the politics of agency*. New York: Oxford University Press.
Skinner, Q. 2006. A third concept of liberty. In *Contemporary political philosophy: An anthology*, ed. R.E. Goodin and P. Pettit, 415–415. New York: Blackwell.
van Hees, M., and K. Dowding. 2009. Freedom of choice. In *The handbook of rational and social choice*, ed. P. Anand, P. Pattanaik, and C. Puppe, 374–392. Oxford: Oxford University Press.

Chapter 7
A Neo-Aristotelian Theory of Individual Liberty

1 Introduction

In this chapter, I develop my own account of negative freedom, focusing on four aspects: the agent's effective freedom, as represented by the extent of his capability set; the agent's republican freedom, understood as his degree of immunity from being deprived of his particular freedoms; the agent's autonomy-freedom (his freedom to develop and exercise the capacity of autonomy); and the extent of the diversity of choice present in the agent's set of available functionings. I draw these elements together into a single formal framework, making use of the pathbreaking advances of a number of social choice theorists. In the conclusion, I unite my account of individual freedom with the account of autonomy I developed in Chaps. 3 and 4, and produce a complete account of individual liberty. This sets the stage for the rest of the book, which argues that it is the allocation of individual liberty, as defined at the end of this chapter, which is the appropriate object of the State's distributive concern and redistributive efforts.

2 Elements of the Right Account of Individual Freedom

2.1 Preference

The idea that preference has some role to play in determining both the extent and the value of someone's freedom is a familiar one. Amartya Sen observes that it is at least counterintuitive that a person whose only options are misery and super-misery is just as free as a person whose only options are happiness and super-happiness, despite the fact that both of these individuals have exactly two options (Sen 2002, p. 600). Richard Arneson points out that Berlin considered the extent of a person's freedom to depend, among other factors, on the importance of each available option to the individual's plan of life (Arneson 1985). And Arneson himself argues both

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_7

that "the individuation of options is relative to what matters to us," and that an individual's freedom increases with an increase in his *vital* options—those options whose very availability results in a preference for them (Arneson 1985, 427).

I do not deny that there are perfectly coherent conceptions of freedom which leave no role for the agent's preferences to play (Pattanaik and Xu 1998). But we are after a particular sort of conception of freedom, one which is fit to play the role of a guiding moral and political value. And it seems clear that a conception of freedom which would consider the two individual's in Sen's example equally free is not the sort we are looking for. But we encounter no shortage of problems when we try to find a way to incorporate preference into our conception of freedom. We certainly do not want to endorse what Arneson calls "the desire thesis" (and which we might rename "the preference thesis"): that the extent of an individual's freedom varies directly with the extent to which his preferences are satisfiable under the options available to him. A few considerations suffice to show why this thesis is stronger than anything we want to endorse.

First, there are the concerns we inherit from Berlin: we must avoid justifying the retreat to the inner citadel, and endorsing tyranny in the name of freedom (Berlin 1969/2006). If we interpret "preference" as used in the above thesis to mean any preference whatever, regardless of how it was formed, then that thesis becomes compatible with precisely the scenario Berlin warned us of. In Sec. 2 of Chap. 6, we found that we could value self-rule without condemning ourselves to this ignoble end, so long as we understood self-rule in terms of exercising the capacity of autonomy as I have defined it. We might think that this result points the way toward an acceptable revision of the preference thesis: that the extent of one's freedom varies directly with the extent to which one's *deliberative* preferences are satisfiable under the options available. But this will not do either. The extent of one's freedom cannot depend on the extent to which the options available to one happen to coincide with the preferences one actually has, even if these are deliberative preferences. The individual who finds himself with a broad range of valuable options has no grounds to complain that he is less *free* than another whose range of options is no greater, just because the latter finds himself with more of the options he most prefers—he is, if anything, merely less fortunate. What role, then, should preference play in our account of freedom? For the answer to this question, we must examine the next element which the account must incorporate.

2.2 Effective Freedom, Capabilities, and Self-Control

2.2.1 Effective Freedom and Capabilities for Functioning

In discussing the distinction between freedom and autonomy, we saw that one important aspect of an individual's freedom is what is referred to as *effective freedom*. An individual has the effective freedom to do something or be some way when he has the resources necessary to do or be it *and* the abilities required to transform

those resources into performing that action or attaining that way of being (and so is free *from* the lack of those resources and abilities). We can begin to elaborate on the idea of effective freedom by introducing the concept of a *functioning*. A functioning is a pattern of actions or ways of being. I assume that there are functionings which are genuinely valuable, and those which are not. Substantive criteria for identifying valuable functionings, however, are unnecessary; procedural criteria will suffice. We may characterize valuable functionings as ones which would be chosen by agents as the result of a well-executed course of ends-deliberation. This procedural method of identifying valuable functionings is only made possible by the thorough and precise account of ends-deliberation developed in Part I. It is the various forms of evidence which are taken into account in the course of such deliberation that support the claim that the functionings chosen through this process are actually valuable. Characterizing valuable functionings in this way allows us to maintain the liberal commitment to competitive value pluralism. Different agents in different circumstances will find different functionings to be most valuable, since much of the evidence that guides these conclusions is position-dependent—the evidence takes into account the importance of various factors which are particular to each individual agent. The way we characterize functionings themselves must be fairly broad—we must include, for instance, reference to the sort of environment in which a pattern of actions or ways of being is pursued as a component of that functioning. In some (rather dire) circumstances, for instance, success at a life of petty theft will count as a valuable functioning; but it will not so count in all circumstances.

Antonio Romero-Medina has developed an axiomatic approach to characterizing the extent of effective freedom offered to an agent by a set of options (Romero-Medina 2001). I will discuss the formal aspect of Romero-Medina's proposal below; for now, the important point is that his method ranks sets of options according to how many valuable functionings each set of options offers, and he identifies valuable functionings as those a reasonable person might choose (though without, of course, offering the sort of account I have developed of the process of arriving at such choices) (Romero-Medina 2001, p. 180). Romero-Medina's account thus puts us on solid ground in adopting the notion of effective freedom, understood in terms of the valuable functionings available to an agent, with procedural criteria for identifying valuable functionings provided by my account of ends-deliberation. Romero-Medina's result allows us to compare sets of options from precisely this perspective and guarantees that we will end up with a ranking of opportunity sets that is complete, transitive and reflexive. This is extremely important because without this result, we could not be sure that in embracing the concept of effective freedom, we were working with a conception of freedom which would allow us to say, of any two opportunity sets, how the extent of freedom offered by one compares with that offered by the other.

We can now answer the question which the discussion of preference left off with. By focusing on effective freedom, we focus on the presence of valuable functionings in an agent's opportunity set, and we understand these functionings as ones which could be chosen as a result of a well-executed course of ends-deliberation. It is fair, then, to describe valuable functionings as patterns of actions and ways of

being that a rational and reasonable agent might prefer. We thus avoid the counter-intuitive consequences of characterizing the extent of an individual's freedom in a way that pays no attention to preference (we no longer need deem the choice between misery and super-misery as offering the same degree of freedom as the choice between happiness and super-happiness), and fortify our view of freedom from the threats described by Berlin, while at the same time avoiding the problem of rich tastes. A functioning can count as valuable, and thus make a difference to the extent of an individual's freedom, even if it is not one the individual actually most prefers (whether deliberately or not). It need only be a functioning which could be deliberatively chosen by an individual in those or similar circumstances—circumstances which will themselves be referred to in the description of the functioning. Crucial to this suggestion is that the process of deliberation is not a strictly deterministic one; given a particular body of evidence, there are a number of sets of conclusions about what to value which can be reached by courses of well-executed ends-deliberation.

Effective freedom, then, seems a promising candidate for the conception of negative freedom we are searching for. But important issues remain unresolved. First, what exactly does it mean to have a valuable functioning as an available option? And second, what about the other important aspects of freedom which we uncovered in the last chapter—republican freedom and the freedom to develop and exercise the capacity of autonomy? I devote the remainder of my discussion of effective freedom to answering the first of these questions. The second will be answered in the sections below.

For the answer to this first question, we must turn to the work of Amartya Sen. Sen begins by discussing commodities—the resources which are required for any given activity or to sustain any given way of being (Sen 1999). Commodities all have various properties, or characteristics. Food, for example, is a commodity which may have the characteristics of being hunger-satisfying, nourishing, gastronomically pleasant, etc. Sen understands functionings as patterns of use of the characteristics of commodities (Sen 1999, p. 7). A functioning—a pattern of action or way of being—is thus identified with the use to which an individual puts the characteristics of commodities necessary to that pattern. To make use of the characteristics of certain commodities in one way rather than another is to achieve one type of functioning rather than another.

Sen then defines what it means for a functioning to be open to a person in terms of the *capabilities* possessed by that person, and identifies the effective freedom of an individual with that individual's capability set. Essentially, a capability is an ability to transform the characteristics of a given set of commodities which are necessary to a functioning into that functioning—an ability to use one's resources to successfully perform activities and initiate or sustain ways of being. We can formulate the notion of a capability set more precisely with the following terminology (Sen 1999, pp. 17–20). Let x_i be a commodity vector, i.e. a particular bundle of commodities, and let X_i be a set of such vectors. Let $c(\cdot)$ be a function that converts actual commodities into their characteristics, and let $f(\cdot)$ be a function which yields one pattern of use (one functioning) to which an agent can put the commodities

accessible to him and F_i be a set of such functionings. Define a way of being of an agent as: $b_i = f_i(c(x_i))$. A way of being is thus the achievement of a functioning chosen by an agent, where that functioning is a pattern of use of the characteristics of the commodities in some commodity vector accessible to him. Suppose an agent has access to any of the commodity vectors in the set X_i. We can then define the set of feasible functionings for that agent: $Q_i(X_i) = \{b_i \mid b_i = f_i(c(x_i))$ *for some* $f_i \in F_i$ *and for some* $x_i \in X_i\}$. This is the set of functionings which the agent can actually attain given the commodity vectors accessible to him. Q_i is therefore this agent's capability set—the set of functionings that are open to this agent. With the notion of a capability set thus defined, we can integrate it into Romero-Medina's model of the extent of an agent's effective freedom. For a valuable functioning to belong to an agent's opportunity set, and thus count towards the extent of the agent's effective freedom, is for that functioning to be a member of the agent's capability set. This means that the agent has both the resources necessary to achieve that functioning and the ability to convert those resources into the functioning itself. Sen's account is incomplete in one respect. We should only count an agent as having a given capability for functioning if that agent's society recognizes his legal freedom to exercise that capability. Moreover, if opportunities to exercise that capability are not scarce and subject to competition, a right against interference with that exercise, protected by an immunity, must be recognized as well. If these opportunities are scarce and subject to competition, a freedom to compete, and a right to non-interference with one's competitive efforts (which are restricted by the similar rights of others), protected by an immunity, must be recognized. We can consider these legal rights to be included among the resources required to achieve a functioning.

2.2.2 Self-Control and Weakness of Will

One functioning that has as good a claim as any to being universally valuable—in fact, as being necessary for the successful achievement of practically any other valuable functioning—is the exercise of self-control. We should understand self-controlled individuals to be those who characteristically stick with and carry out the intentions they form on the basis of their value judgments, even in the face of strong temptation to do otherwise. The phenomenon of self-control is an increasingly popular object of study for empirical psychology. Self-control does seem to be the exercise of a capacity whose development and exercise can be furthered or hindered by a range of factors, and it does seem to be within the scope of the average individual's abilities to increase his level of self-control, and broaden the range of circumstances in which he successfully exercises it. The exercise of self-control is inhibited when one's level of available energy for neural/mental activity is low, a phenomenon known as "ego-depletion." The structure, order, and number of choice situations one encounters have a significant effect on the extent of ego-depletion one experiences. Self-control can thus be enhanced through the social engineering of commonly

encountered choice situations as well as through the effort of the agent.[1] The freedom to exercise one's capacity for self-control is thus an essential part of an agent's effective freedom.

Given what we know about self-control, it is clear that its exercise is a valuable human functioning in precisely Sen's sense of this term. Whether one has the freedom (the capability) to develop and exercise self-control is a matter of whether the individual (a) has access to the necessary resources—such as proper nutrition, which is needed to delay the onset of ego-depletion—and (b) is situated in a social context in which he stands a reasonable chance of succeeding in his efforts to exercise it. Insofar as the capacity for self-control—along with the capacity for means- and ends-deliberation—is a dimension of the capacity of autonomy, this point forges a strong connection between our notions of autonomy and freedom.[2] One very important part of effective freedom is thus the effective freedom to develop and exercise the self-control dimension of autonomy. The rational aspect of autonomy—deliberation about ends and means—is just as obviously another valuable human functioning. Good instrumental deliberation, like self-control, is necessary for achieving practically any other functioning. And so the effective freedom to develop and exercise the basic rational capacities required for excellent deliberation about ends and means is another important part of effective freedom. I discuss the relationship between effective freedom and the rational dimension of autonomy in greater detail in Sect. 2.4 below. But before we move on from the topic of self-control, we should ask whether the decision-theoretic background I have employed in representing the rational aspect of autonomy can be used to illuminate the exercise of this other dimension of autonomy as well. This would provide us with a single, unified, precise account of autonomy.

In order to represent the exercise of self-control, we need to understand the phenomenon which that exercise overcomes—weakness of will. There is a vast philosophical literature, stretching back to the ancient Greeks, on the topic of weakness of will. But fortunately, nearly all of it is irrelevant to our current purposes. The standard view of weakness of will is expressed in the slogan that 'weak-willed action is action against the agent's better judgment'. But the view of this phenomenon which I accept—the one I believe is correct and the one which I will go on to model—is outside the standard view. I will therefore provide only a brief discussion of the contemporary apex of the standard view, as developed by Donald Davidson, and give my own reason for rejecting it. I will then present an alternative view, drawing in part on recent work by Richard Holton.

Davidson endorses two basic claims. First, that an agent free to choose either of two options will choose the one he wants more; and second, that an agent who judges one option better than another will want to choose the option judged better more than the option judged worse (Davidson 1980). Davidson's insistence that weak-willed actions—which he understands as actions that go against the agent's better judgment—are possible relies on a distinction between two types of

[1] The work of psychologist Roy Baumeister is particularly relevant to all these points (Baumeister et al. 1998, 2005, 2006, 2008; Baumeister 2002; Baumeister and Vohs 2007; Vohs et al. 2012).

[2] See the Introduction to Part I.

judgments: all-things-considered judgments, and all-out judgments. An all-things-considered judgment that *a* is better than *b* has the form:

Given all the available evidence $\boldsymbol{E} = \{E_1 \& E_2 \& \ldots \& E_n\}$, *a* is better than *b*.

The corresponding all-out judgment has the simpler form:

a is better than *b*.

Davidson's view is essentially that a weak-willed action to choose *b* is a choice of *b* despite an all-things-considered judgment that *a* is better. He maintains that such weakness is impossible in the face of an all-out judgment that *a* is better, since an agent who makes such a judgment will want to choose *a* more than *b*.

The main problem with Davidson's view as he states it is that the crucial distinction between all-things-considered and all-out judgments does not hold up. This is most easily seen by representing these judgments in a Bayesian framework. The proposition that *a* is better than *b* is a hypothesis *H*: $a \succ b$. Like any hypothesis, the agent attaches a subjective probability to its truth: $p(a \succ b)$. To make an all-out judgment that *a* is better than *b* is to assign a probability to $a \succ b$ which is greater than the one assigned to $b \succ a$: $p(a \succ b) > p(b \succ a)$. The agent's credence in the truth of $a \succ b$ given some body of evidence $\boldsymbol{E}$ is $p(a \succ b \mid \boldsymbol{E})$. This seems to match Davidson's characterization of an all-things-considered judgment. But Davidson understands such judgments as based on *observed* evidence. So suppose the agent believes he has observed this body of evidence $\boldsymbol{E}$. He will now attach a probability to the truth of $\boldsymbol{E}$: $p_{new}(\boldsymbol{E})$. The existence of these latter two judgments entail an updated judgment $p_{new}(a \succ b) = p(a \succ b \mid \boldsymbol{E}) \cdot p_{new}(\boldsymbol{E}) + p(a \succ b \mid \text{-}\boldsymbol{E}) \cdot p_{new}(\text{-}\boldsymbol{E})$. And likewise for updating $p(b \succ a)$ to $p_{new}(b \succ a)$. If $\boldsymbol{E}$ confirms the hypothesis $a \succ b$, then we will have $p_{new}(a \succ b) > p_{new}(b \succ a)$. And here is the problem with Davidson's view: an all-things-considered judgment, based on observed evidence, necessarily translates, via Bayesian updating, into an all-out judgment.

But perhaps Davidson's focus, on what he sees as the difference in form between these two ways of expressing a judgment, is simply misplaced given the point he is likely trying to make. Perhaps his point is better made by focusing on the distinction between judgments made before an agent reaches a deliberational equilibrium, and those made once the agent has reached a deliberational equilibrium. *Prima facie* judgments would correspond to the former, all out judgments to the latter. So let us reexamine Davidson's claim in the context of Skyrms' dynamic model of instrumental deliberation. Let us conceive of coming to a conclusion about whether a given option is best as an action. The agent is then deliberating about whether to come to the conclusion that option *a* is best, or that *b* is. Each action has two possible outcomes: concluding that *a* is best either rightly or wrongly, and likewise for *b*. Let us assume that the values of concluding either rightly, or either wrongly, are equal, with concluding rightly obviously valued above concluding wrongly—the agent is not biased at the outset, he simply wants to arrive at a responsible judgment. The expected values need not be equal—if the agent thinks it more probable that *a* is best, based on the evidence he has seen so far which seems to support that conclusion, then the expected value $V(a)$ will be greater than $V(b)$. Despite assigning

a higher expected value to concluding that *a* is best right now, the agent has not yet committed to that conclusion—he is still in a state of indecision. The value of the state of indecision, call it *V(q)*, is the average of these. The agent then moves, *via* deliberation, from one state of indecision to another until he eventually reaches a deliberational equilibrium, in precisely the way described in Chap. 4. Once he reaches that equilibrium, whichever option has the highest expected value is the option he finally concludes is best. Up until that point, all judgments about which option was better were *prima facie*. But at the equilibrium—and in virtue of having reached the equilibrium—the agent makes an all out judgment that one is better than the other.

It seems to me that the idea of a judgment made once the agent has reached a deliberational equilibrium is much closer to the idea of an all out judgment that Davidson wants than is anything he says about the logical form of such judgments. But although we may have illuminated Davidson's discussion, it is not at all clear that we have helped his cause. The idea that weakness of will is impossible once a deliberational equilibrium regarding what conclusion to come to has been reached is scarcely credible. As Michael Bratman has pointed out, there is nothing strange about the possibility of having reached a settled conclusion about what is best—in which one judges one option better than another, is no longer deliberating, does not expect that any more evidence which will alter one's judgment is readily forthcoming, and is comfortable reflectively endorsing the judgment—and yet failing to do what one judges best all the same (Bratman 1979). This is a perfectly mundane aspect of ordinary human experience.

The only recourse for Davidson's view is to maintain that an agent only makes an all-out judgment when he makes a judgment that he considers to be 'un-updatable': a judgment with respect to which the agent is certain that there is no further evidence to be had. The fact that an agent makes a judgment having reached a deliberational equilibrium does not, of course, mean that he thinks the judgment is un-updatable. An unforeseen experience of the world, occurring after deliberational equilibrium but before he has a chance to act (or to complete a course of action), may upset his conclusions and throw him back into deliberation—and any reasonably circumspect agent will recognize this as a possibility. For an agent to see a judgment as un-updatable would amount to the agent reaching what he takes to be a permanent and unalterable conclusion that *a* is better than *b*. But this is hardly a move Davidson would want to make. Such judgments are sure to be exceptionally rare, and thus, Davidson's attempt to remain committed to even a mild form of judgment internalism would lose all of its bite.

Now, perhaps we don't care about judgment internalism. In fact, given the model of deliberation, decision, intention and action developed in Chaps. 3 and 4, judgment internalism seems obviously false. If we identify an agent's "better judgment" with the preferences the agent is most confident he ought to have—the preference-ranking to which he assigns the highest probability—it is clear that an agent can act against his better judgment. If the agent is not what I have characterized as a truly excellent ends-deliberator, and his desirabilities are out of sync with those probability judgments, the agent may decide to adopt a preference-ranking other than the one

recommended by his "better judgment." The agent's valuations of options are based on the preference-ranking he adopts, and his preferences over actions track the expected values calculated using those valuations. So let us suppose we are judgment externalists of some moderate stripe—moderate because we only deny that an agent's "better judgment" necessarily plays a decisive role in the determination of his preferences and his actions, while acknowledging that it often exerts a significant influence. This might seem to take the mystery out of weakness of will. But in fact, the issue only becomes more perplexing. If we are comfortable with moderate externalism, we may be able to understand how it is possible for an agent to act against his better judgment, but we are left wondering why there is supposed to be something inherently wrong about this. After all, those judgments may very well be wrong. By acting against his better judgment, the agent may end up doing what he in fact should do (Arpaly 2000). And yet, there is supposed to be something wrong with weakness of will—we mean to use this term to pick out a type of action that is to be avoided, to refer to actions the agent should not, in some sense, perform. And it will hardly do to restrict the definition, and call an agent weak-willed only when he acts against his better judgment *and* his better judgment happens to be correct. Whatever exactly we mean by weakness of will, whether an agent exhibits it or not cannot turn on whether or not his judgments happen to be right.

The key to progress on this front is to revise our basic concept of weak-willed action—to identify it with something other than action against one's better judgment. Recall the characterization of self-control given above: self-control is sticking to one's prior intentions, despite the temptation to abandon them. Weakness of will is a failure of self-control. A weak-willed agent is one who succumbs to temptation and abandons his prior intentions as the time to act on them approaches, swapping them for other intentions that seem, in one way or another, more expedient. That weakness of will should be understood in this way was first suggested in the philosophical literature by Richard Holton (Holton 1999).[3] Holton is rightly concerned to find criteria that will allow us to distinguish weak-willed action from other (non-weak-willed) changes of mind. He offers two: first, that the agent should not change his intention; and second, that the prior intention is a resolution, i.e. that it is meant to exclude later intention change. The first criterion makes it clear that whenever there is a case of weakness of will, something has gone wrong: the agent has done something he ought not. Holton maintains that the first criterion is "irreducibly normative"—that it cannot be cashed out in terms of any fact about what the agent believes or judges to be best.

Holton shifts the focus of the discussion on weakness of will in the way needed. But there is something deeply unsatisfying about the idea that, although weak-willed action results from a change of intention which the agent should not make, there is simply nothing to be said about what makes it the case that the agent should not change his intention. What is needed is an account of intention-change that

[3] Holton allows that there is something distinct from weakness of will, which he refers to by the Greek term *akrasia*, which is action against one's better judgment.

makes it clear which cases are illegitimate and why they are so. The beginnings of such an account can be found in two remarks of Alfred Mele (despite the fact that he is a defender of the standard view, which we are trying to move away from). The first is that an agent's wants and attachments may be well out of line with the agent's judgments concerning the value of what he wants or is attached to (Mele 1987, ch. 1). This fact, as we have already seen, is captured by my model of ends-deliberation.[4] The second is that one of the most prominent factors in driving judgment and attachment apart is temporal proximity (Mele 1987, ch. 6).

The first type of case we need to deal with is that of weakness of will in the face of simultaneous choice. This is the normal choice situation, in which two or more options are available at the same time and the agent must decide between them. For simplicity's sake, let us suppose the agent only has two actions to choose from, and he knows which outcome each will lead to—he is making a choice under certainty. Suppose that before he makes his choice—even before the two actions become available—the agent prefers action *a* to action *b*, and so decides he will do *a* and intends to do so. The expected values in this case are just equal to the agent's valuations of the outcomes, which reflect the position of the outcomes in the agent's preference-ranking. Suppose further that the agent's preference-ranking is the one he judges most likely to be correct. As the time for his decision draws near, and the agent thinks about outcome *b*, the desirability of the preference-ranking that ranks outcome *b* above outcome *a* begins to increase. Shortly before the time to act, that desirability has increased sufficiently to change the agent's choice of what preference-ranking to adopt. He thus re-values the outcomes of his upcoming actions, with *b* now valued more highly than *a*, and his intention changes accordingly. He chooses *b*. (To model this process endogenously, a model analogous to the feature-based preference model of Dietrich and List discussed in Chap. 3, with its focus on attention and motivational salience, would be useful.) A closely related case is one in which the agent has a deliberative preference for *a* over *b*, but an affective preference for *b* over *a*—*b* is ranked above *a* in the preference-ranking with the greatest desirability. As the moment of choice approaches, the agent's deliberative ranking is replaced by his affective ranking, and his intention and choice change accordingly.[5]

[4] It is important to note that my formal framework itself does not simply beg the question against those who affirm that evaluative judgment dictates intention and action. One who held this view could perfectly well use my framework, and simply insist that desirability assignments always track probabilistic evaluative judgments, and represent them as such. But in representing deliberation about the adoption of preferences as I do, I succeed in avoiding both this strong variety of internalism about the relationship between judgment and action, and a strong externalism which denies that judgment has any influence over preference and choice.

[5] This seems to be the sort of case that Aristotle identified as the paradigm case of *akrasia*: the akratic temporarily loses his grip on what he knows is valuable, as one who is drunk or in a stupor loses his understanding of things he knows while sober (Aristotle *NE* VII.3 1147a10-22). This is also the sort of case alluded to in Chap. 2. Aristotle's other case of *akrasia* in that chapter is the impulsive person who simply acts on his appetites without deliberating at all, in a way which does not reflect the conclusion he would have come to if he had deliberated.

These are cases of weakness of will. The agent changed his intention based on a change in the way he valued the outcomes of his available actions. But—and this is the crucial point—he made that change based on an unreasoned change of mind about what preference-ranking to adopt. This latter change was not based on any newly acquired reason or evidence; nor was it based on a taste change. The change is simply caused by the agent's perception of the temporal approach of outcome *b*. The agent regrets the choice of *b* practically as soon as he has made it, and reverts to a preference for *a* over *b*, to be acted on at the next opportunity (*NE* 1150b29). He has failed to exercise self-control, one important manifestation of which is precisely preventing intention changes of this type from happening. By failing in this way, the agent has put to waste any time and effort he expended on deliberating about what preferences to adopt in the first place. The action is weak-willed whether the agent ends up doing what he actually ought to do or not. And even if the agent does end up doing what he ought to do, it is not the case that he ought to have changed his intention. He may have ended up with the intention he ought to have, but this is a different point. If the agent's judgments about his potential preference-rankings were wrong, then what he ought to have done is continued to seek out evidence and revised these judgments, and so ultimately revised his choice of preference-ranking.[6] Then, on the basis of that revised choice of preference-ranking, he ought to have decided what to do, and stuck to his intention.[7]

This case of weakness of will is a case of action against better judgment. But we can immediately see that there is a companion case of action against better judgment that is not a case of weakness of will. We need only suppose that the agent, after imperfect deliberation, adopts a preference-ranking other than the one he judges is most probably the one he ought to adopt. The agent sticks to his intention, and chooses the option he values most highly, based on the preference-ranking he has actually adopted. This is a case of action against better judgment, but not a case of weakness of will. The agent has deliberated, intentionally adopted a set of preferences, decided what to do based on those preferences, and stuck to that decision all the way through. There is something different which is wrong with what this agent has done. Obviously, he has not deliberated over his preferences as well as he might have. If his judgments are right, then what this agent ought to do is cultivate emotional attachments that will allow him to adopt the preference-ranking he ought to adopt, and then form, and stick to, intentions to act on the basis of that preference-ranking. And if his judgments are wrong, he ought to revise them.

There are other cases in which the agent goes wrong in progressing from deliberation through decision and intention to action, but which are neither cases of acting against better judgment nor of weakness of will. Suppose an agent has an intention to perform action *a* rather than action *b* at some time in the future. His preference for *a*-ing is in line with his better judgment. Before the time for action

[6] I suspect that most, if not all, of Nomy Arpaly's examples of rational/reasonable weakness of will involve agents who have previously exercised poor ends-deliberation (and thus, even these cases have irrationality at their roots).

[7] I am of course assuming throughout that normative requirements are wide-scope 'oughts.'

arrives, however, the agent comes across what he takes to be evidence that he ought to prefer the outcome of *b* to the outcome of *a*. In light of this evidence, he updates his probability judgments, and that is sufficient for him to adopt a different preference-ranking, in which the outcome of *b* is preferred to the outcome of *a*. He changes his intention accordingly, and when the time comes, he does *b*. But suppose that he was right the first time, and this so-called evidence in light of which he changed his mind was not credible—it misled him. In the end he fails to do what he actually ought to have done, which is action *a*. This is clearly not a case of action against better judgment, since at the time the agent acted, he was acting in accordance with his better judgment at that time. Nor is it a case of weakness of will.[8] Given the magnitude of the change in his probability judgments, it is not the case that the agent should not have changed his intention. His intention change was the appropriate response to his revision of his judgments. The agent, in performing action *a*, did not do what he ought to have done—but it is *not* the case that what he ought to have done was stick with his intention to do *b* despite changing his judgment regarding which preference-ranking he ought to have. Rather, what he ought to have done is to have been more skeptical of the evidence that ultimately misled him, to have searched harder for additional reasons to either change his preference-ranking or keep it the same. And after finding additional evidence, he ought to have changed his mind back, reverted to his original intention, and then stuck with it.

So here we have one part of our characterization of weakness of will: weak-willed actions are action that result from a change in intention, which change in intention results in turn from an unreasoned inflation of the desirability of a temporally proximate option.[9]

The next type of case we must examine is one of non-simultaneous choice, in which an agent could perform one action at one time, or forego it and then perform a different action at a later time. The key to understanding weakness of will in these cases is to understand temporal discounting. There are two main ways in which an agent can discount future value. The first is *exponential*.[10] Suppose an agent has a choice between two options. One, *l*, can only be attained at time t_l, and the other, *s*, can only be attained at time t_s, $t_l > t_s$. Suppose that he prefers *l* to *s* (his

[8] This scenario strikes me as being reasonably close to the one described by the Stoics, who did not believe in genuine weakness of will, as being what actually happens in all cases commonly thought of as weakness of will: the agent changes his mind very quickly, performs the action, and then changes his mind back just as quickly, and regrets the action. The Stoics taught that this could happen so quickly, the agent would not even notice it, and so would mistakenly believe of himself that he had done something he thought was contrary to his judgment. There may, however, be a case of genuine weakness of will parallel to the one described here. If it is possible for the agent to be *merely caused* to change his judgment (in the absence of any response to apparent evidence, and in the way that he may be caused to change his desirability assignments in the first case), and thereby change his intention, that would count as a case of weakness of will and as a case of action against one's better (pre-causal interference) judgment.

[9] Or an unreasoned, merely caused change in judgment about what is valuable, if we accept this possibility. See the previous note.

[10] The following discussion of discounting is based on (Moldoveanu 2011, pp. 51–53).

preference-ranking R is one in which l is ranked higher than s), and thus he values l more highly than s: $V_l > V_s$. This is to say that if both options were going to become available at the same time, the agent would plan to (and ultimately would) choose l. We are interested in what happens to the agent's valuations as the time when s is within reach—but l is still unattainable—approaches. As an exponential discounter, the agent's valuation of s at a time t, $v_s(t)$, will equal $V_s \cdot e^{-k(ts-t)}$ for some positive constant k which is the agent's discount rate. At $t = t_s$, then $v_s(t) = V_s$. Likewise for $v_l(t)$. I have no intention to argue that temporal discounting is inherently irrational. I am willing to concede that exponential discounting may in fact be perfectly rational.[11] What matters for us is that an agent who discounts exponentially will never invert his valuations: at no point in time prior to t_s will it become the case that $v_l(t) < v_s(t)$ (Moldoveanu 2011, p. 52). This is not the case, however, for an agent who discounts in the second way: *hyperbolically*. Such an agent discounts thus: $v_s(t) = V_s/1 + k(t_s - t)$, and likewise for $v_l(t)$. For a hyperbolic discounter, there may be a point in time prior to t_s at which it becomes the case that $v_l(t) < v_s(t)$ (Moldoveanu 2011, p. 52). If there is, then at that point in time, assuming that the agent intends to choose the option which he values most highly at any given point in time, the agent's intention changes. He then intends to choose s rather than l, and will continue to do so until t_s arrives. He will then choose s, forgoing l.

Under what conditions will there be such a point in time, call it t_c, when the agent abandons his intention to choose l, and decides he will choose s instead? Whenever all of the following three conditions are met:

(1) The difference V_l ‘V_s is *insufficiently large*.
(2) The difference t_l ‘t_s is *insufficiently small*.
(3) The difference t_s ‘t_0 (the latter being the time the agent first learns he has to make the choice) is *insufficiently small*.

Each of these conditions needs to be explicated. I will do so by working through a series of related examples. The point t_c is the point at which $v_s(t) = v_l(t)$. We solve for t_c as follows:

$$t_c = 1/k + (V_s)(t_l) - (V_l)(t_s)/V_s - V_l$$

The constant k is the agent's discount rate. The larger it is, the less the agent cares about the future. Let us choose a relatively small value, and set $k = 2$. Suppose the agent must make a choice between receiving \$60 in 60 min and receiving \$75 in 75 min. So let's set $t_0 = 0$, $t_s = 60$, and $t_l = 75$. Assume for simplicity that $V_{\$60} = 60$ and $V_{\$75} = 75$. This agent will intend to hold out for the larger payout at t_0, but abandon that intention at $t_c = .5$(i.e. after 30 sec). Now, let us suppose that l is significantly larger—say \$1000. Because the agent is a hyperbolic discounter, he will still end up abandoning his intention and choosing \$60 after 60 min. It will just take him longer to do so—in this case, 59.54 min. Raise l to \$1800, and the agent still takes the \$60 payout—changing his intention at 59.98 min. The agent will not be able to

[11] For a defense of the rationality of discounting, see (Heath 2008, pp. 234–241).

hold out for the larger payout until the value of l reaches \$1860.01. A hyperbolic discounter with a relatively low discount rate, then, will not be able to stick to his intention to forego a \$60 payout and wait an additional 15 min, until the later payout exceeds the earlier by over \$1800. Given the discount rate and the time intervals, any difference less than that is insufficiently large to enable him to stick to his initial plan. We can keep $l = \$75$ and achieve the same result (of enabling the agent to stick to his intention to wait for the larger payout), by requiring that he not wait so long to receive it. But in order to do this, we have to set $t_l = 60$ min and 7 sec. So the hyperbolic discounter will not be able to stick to his plan to forego \$60 for the sake of an additional \$15, unless we only require him to wait no more than seven additional seconds. Given the discount rates and the payouts, any difference less than this is insufficiently small to enable him to stick to his initial plan. And again, the low value of k tells us that this is a hyperbolic discounter who cares quite a bit about the future.

Hyperbolic discounters, then, are able to stick to their initial intentions to wait for better outcomes in what I will call extreme choice situations—ones in which the later payout is very large relative to the earlier one, or in which the wait between the earlier and later payouts is very short relative to the wait between the time at which the initial intention is formed and the time for the earlier payout. There is one other way for hyperbolic discounters to avoid abandoning an earlier intention to wait for a larger payout: by lack of forethought. Suppose the agent is choosing between \$60 at 1:00 and \$1000 at 1:15. If the agent can avoid forming any intention at all until 12:59:33—a mere 27 sec before the first payout becomes available—he will simply prefer to, and thus intend to, choose the earlier payout rather than the later. This too is an intention he will be able to stick with. But given the discount rate, the payouts, and the time interval between payouts, any difference between the time of initial intention and the time of early payout greater than this is insufficiently small. The agent will start out intending to wait, and then abandon this intention.

In cases of non-simultaneous choice, then, I define the weak-willed agent as one who performs an action based on a change in intention which in turn results from hyperbolic discounting of future value.[12] We can now see what is wrong with being weak-willed in cases of non-simultaneous choice. It is surely uncontroversial that there are (a) non-extreme choice situations in which an agent should be able to (b) form an initial intention to wait for a better outcome, with (c) a reasonable amount of forethought, and (d) stick to that intention. Hyperbolic discounting makes this impossible. Therefore, we should not discount future value hyperbolically, and so we should not be weak-willed in cases of non-simultaneous choice.

[12] Aristotle also seems to recognize this type of case of weakness of will: "[A]ppetites run counter to one another, which happens when a principle of reason and a desire are contrary and is possible only in beings with a sense of time (for while thought bids us hold back because of what is future, desire is influenced by what is just at hand: a pleasant object which is just at hand presents itself as both pleasant and good, without condition in either case, because of want of foresight into what is farther away in time)…" (Aristotle *De Anima* III.10 433b5-11 in Barnes, ed. 1984).

If we identify the agent's valuations V_l and V_s with his "better judgment," then cases of weak-willed action in the face of non-simultaneous choice will also be cases of action against better judgment. And so weak-willed action will turn out to be action against better judgment after all. But we will still not be able to identify weakness of will with action against better judgment, since as we have seen, there are cases of action against better judgment which are not cases of weakness of will. There is a third type of weakness of will, which is also action against better judgment, but which is not captured by Holton's definition. This is the case of failure to change one's intention when one believes that one ought to. I hold off discussion of this case until Chap. 14, when I will discuss ethical deliberation. But we will see that it can be explained in a way that mirrors the explanation of the first kind of case. For now, we can summarize the position developed so far as follows:

(WW): Weak-willed actions are those actions (1) that result from a change in intention which is due to either (a) the unreasoned inflation of the desirability of certain kinds of options as they draw temporally near, or (b) discounting future value hyperbolically; or (2) that are performed after a failure to change an intention despite the judgment that it should be changed.

Despite departing from Holton's first criterion in some important ways, then, we have retained his central point, which is that weakness of will is essentially a matter of whether or not an agent does, and should, change his intentions.

Holton's second criterion for distinguishing weakness of will from non-weak-willed changes of mind is also defective. To view one's forming an intention as a resolution—as something meant to exclude later changes of mind—is simply one example among many of a strategy for self-control.[13] Other such strategies include (but are not limited to) the development of habits and routines; exercises in attention control; creating obstacles that prevent one from being exposed to temptation; and purposefully exposing oneself to temptation in a controlled way to enhance one's ability to resist it. Self-control is what the weak-willed agent lacks. So we may now interpret such strategies as attempts by the agent to avoid either (a) inflating the desirability of certain kinds of options as they draw temporally near, or (b) discounting future value hyperbolically. For now, this is our official characterization of the exercise of self-control: it is the employment of a strategy for avoiding (a) or (b). (In Chap. 14, I will cast certain failures to change one's intentions as failures of self-control as well.) The agent who is suffering from ego-depletion is thereby impaired in his ability to employ such strategies. But Holton's mistake is not simply a lack of generality. An agent may be weak-willed even if he does not employ any strategy to actively combat his tendency toward inflating the desirability of certain kinds of options, or to discounting the future hyperbolically. So long as he knows that he has these tendencies, he shows his weakness in his lack of effort at combating them, just as he would in a failed effort to combat them. So the above characterization of

[13] For a list and discussion of different categories of self-control strategies, see (Moldoveanu 2011, pp. 70–75; Heath 2008, pp. 246–254).

weak-willed action—as action that results from intention change which in turn results from desirability inflating or hyperbolic discounting—stands.

Alison MacIntyre identifies a potential problem in Holton's account of weakness of will—one which I am quite concerned to avoid (MacIntyre 2006). By severing the tie between weak-willed action and the agent's better judgment, Holton makes it difficult to see why weak-willed actions are supposed to be distinctively irrational. My account of weakness of will does not really sever this tie in the way that Holton's does—I allow that all weak-willed actions are actions against better judgment without identifying, or asserting an equivalence of, weak-willed action with action against better judgment. But nonetheless, I can gladly accommodate the suggestion MacIntyre makes for preserving the connection between weakness of will and irrationality in the context of Holton's account, and frame it as an additional reason to count weak-willed action as irrational. MacIntyre suggests that an agent who has formed a resolution but fails to carry it out is procedurally irrational. This point generalizes to all types of strategies of self-control. But, in the context of our account of weakness of will, it also applies to agents who perform weak-willed actions without having adopted any such strategy. In these cases, the procedural irrationality is in knowing that one is prone to inflate the desirability of certain kinds of options, or to discount the future hyperbolically, and yet failing to do anything to combat that propensity. Moreover, we can now also see that weak-willed actions are irrational, insofar as allowing oneself to act on these propensities is irrational—which it surely is. The ability to plan for the future, and to stick to one's plans in situations in which a weak-willed agent cannot, is an essential part of being practically rational.

The last point I wish to make about this way of modeling weakness of will/self-control is that it gives us a basis for holding agents responsible for their blameworthy weak-willed actions when it is appropriate to do so, but also correctly identifies those cases in which it is not appropriate. One problem with certain versions of the standard view, for example Mele's, is that by attributing weak-willed actions to wants of overwhelming force which run counter to the agent's better judgments, they alienate the weak-willed agent from his motivation, to the point where it scarcely makes sense to hold the agent accountable for his action (Tenenbaum 1999). My view allows us to assert that agents who have the effective freedom to develop the capacity for self-control are responsible for developing and exercising it. Since blameworthy weak-willed actions result from a failure to do so, the agents that perform them are responsible for their failure. On the other hand, agents who lack the effective freedom to develop self-control are not responsible for their failure in performing such actions; but nor should they be held accountable, given that they lacked the opportunity to develop into autonomous agents.

Before moving on to discuss the next aspect of freedom—the republican aspect—let us draw some threads together. We identified effective freedom as one important aspect of a suitable conception of individual freedom. Effective freedom is the freedom to pursue and achieve valuable functionings. There were a series of questions which had to be answered in order to ensure that this notion of effective freedom could be made precise. The first was whether this notion of freedom could be characterized in such a way that we would be able to determine, with respect to any

two opportunity sets, whether or not one offered at least as much freedom as the other. Romero-Medina's axiomatic characterization of effective freedom, the technical details of which I discuss below, ensures that we can indeed do this. The next question was how to identify which functionings were to count as valuable. We saw that my account of ends-deliberation is well-suited to provide a precise answer to this question. My account also allowed us to answer the third question, concerning the role of preference in determining the extent of an agent's freedom. Finally, we had to ask what exactly it meant to say that a valuable functioning was an option open to an agent. Here, Amartya Sen's account of freedom as capability provided the answer.

We also identified two universally valuable functionings as engaging in excellent ends-deliberation and exercising self-control, and showed how to model self-control/weakness of will using the apparatus of decision theory. We are thus well on our way to articulating a conception of freedom that satisfies one of this project's basic Aristotelian commitments:

Substantive Agent Freedom: The extent of an agent's freedom is the extent of that agent's capability set—the set of ways of life he has a real opportunity to lead. These ways of life are constituted by the valuable functionings—the valuable states of being, actions and activities, and projects and goals—which the agent has a real opportunity of realizing.

As a bonus, we also connected blameworthy weak-willed acts with failures in ends-deliberation, and completed the presentation of a single, unified, precise account of autonomy in all its dimensions. Let us proceed, then, to incorporate some important additional refinements into the conception of freedom developed thus far.

2.3 Republican Freedom

Republican freedom is the aspect of freedom which the opponents of the unlimited royal prerogative were concerned with. As discussed above, it is freedom from one specific and important constraint: the *lack* of an immunity against another person's arbitrary exercise of power. As Skinner's discussion of republican freedom pointed out, when one lacks this important immunity, there are all sorts of things one is constrained from saying and doing. Since republican freedom is an important aspect of individual freedom, the question we are now faced with is how to integrate this notion with the account of effective freedom just discussed. On this point, we can turn to the work of Phillip Pettit (Pettit 2001). Pettit has helpfully observed that Sen, by characterizing individual freedom in terms of an individual's capability set, has understood possessing the freedom to do something as having a *decisive preference* for doing it (Pettit 2001, p. 2). When one has a functioning in one's capability set, one has a viable opportunity to achieve that functioning. All that one need do in order to pursue that functioning with a realistic chance of success, then, is settle on that functioning as the object of one's preference. Insofar as this is true regardless

of which functionings in one's capability set one prefers, and regardless of what the powers-at-be in one's society would prefer that one do, one's preference is decisive. Pettit calls the first of these two conditions content-independence, and the second context-independence (Pettit 2001, pp. 5–6). I would add that Sen's account does not incorporate full content-independence for preferences, since it judges the extent of the agent's freedom based on the *valuable* functionings present in the agent's capability set. We should say, then, that the capability approach understands having a freedom as having a decisive *deliberative* preference—i.e. a preference for a valuable functioning from one's capability set, one that might result from a well-executed course of ends-deliberation by a similarly situated agent.

If this is the right way to understand possessing a freedom according to the capability approach (and I think it is), then the door is open to incorporating republican freedom into our account of effective freedom. For as Pettit observes, to possess republican freedom is to have preferences which are decisive independent of context, and this is one aspect of having a preference for a functioning that is within one's capability set (Pettit 2001, p. 7). More specifically, republican freedoms are freedoms which are decisive independent of *favoring*. If freedoms are bestowed *via* the exercise of sovereign power, and the sovereign prerogative is unlimited, then to possess a freedom is to be granted a favor which may be revoked at any time. If one were to exercise a freedom in a way that displeased the sovereign power, that freedom could simply be extinguished through the imposition of a new duty. Since, in this situation, one's preference would not be decisive independent of favoring, one's preferred functioning would not belong to one's capability set, and so one would not have the effective freedom to pursue and achieve that functioning. This pursuit could be absolutely and arbitrarily thwarted by the action of another at any time. Republican freedom is thus encompassed by capability-freedom. And as we saw above, Sen's notion of capability is the core of the idea of effective freedom. We can thus see republican freedom as one aspect of effective freedom—an agent who has the effective freedom to pursue and achieve some functioning has the republican freedom to do so as well.

2.4 Autonomy-Freedom

The next important aspect of individual freedom which we must integrate into the conception we are developing is autonomy-freedom. This sort of freedom comes in two varieties. The first is development-of-autonomy freedom. This is the freedom which an individual requires in order to develop the capacity of autonomy. The second is exercise-of-autonomy freedom. This is the freedom to exercise one's developed capacity of autonomy by both forming preferences through ends-deliberation and making choices on the basis of those preferences.[14] In what follows, I will

[14] Thomas Hurka uses the term "deliberated autonomy" to refer to free choice among numerous options that reflects practical reasoning about those options. He identifies this as an Aristotelian conception of autonomy (Hurka 1993, pp. 151–152).

argue that exercise-of-autonomy freedom is a somewhat restricted version of effective freedom, and that the restriction is a welcome one. The conception of freedom we have been searching for will then turn out to be the effective freedom to develop and exercise the capacity of autonomy, where this capacity is understood according to the account I have developed in Part I. Once this has been shown, I will introduce one final element into this conception—the element of diversity of choice.

Sebastiano Bavetta and Francesco Guala have developed an axiomatic characterization of exercise-of-autonomy freedom (Bavetta and Guala 2003). I will argue for the proposed relation between autonomy-freedom and effective freedom by showing that the formal structure of their account is a restricted version of the formal structure of Romero-Medina's account of effective freedom. This will also allow me to fulfill my earlier promise to discuss the formal aspect of Romero-Medina's account. Bavetta and Guala understand the capacity of autonomy in the same way that I do. They accept the claim that the capacity of autonomy is the capacity to deliberate about which preference-ranking to adopt from a range of possible preference-rankings (Bavetta and Guala 2003, p. 432). They are, however, unable to see how this process of deliberating over preference-rankings could be formally represented without smuggling in controversial assumptions about higher-order preferences (Bavetta and Guala 2003, p. 433). My account avoids this problem in virtue of its basic Aristotelian structure. The only assumption made is that the agent wishes to figure out how to lead a worthwhile life. But pre-deliberation, the concept of a worthwhile life is a thin one; no assumptions are made about what leading a worthwhile life amounts to. Rather, the agent fills in a conception of a worthwhile life *through* the process of deliberation itself.

Bavetta and Guala give a fairly straightforward characterization of exercise-of-autonomy freedom. This type of freedom strictly increases as the number of options of which an agent is *aware* increases. They define awareness of an option as knowledge of what it would be like to choose that option (Bavetta and Guala 2003, p. 434). By this they seem to mean something like the ability to successfully imagine what it would be like to choose an option of that kind. They give the example of choosing between seeing an action movie and an *avant-guard* French film. The agent is aware of both of these options so long as he can successfully imagine what it is like to see a film of either type, even though he has not seen either one of these particular films and does not know any specific details about either one (Bavetta and Guala 2003, p. 434). They allow that one's awareness of an option may be based either on one's own past experience of a similar option, or on testimony about options of a certain type (Bavetta and Guala 2003, pp. 435–436). They also allow that testimony may play a role in the deliberative process in which one determines which preferences to adopt—as it does in my account of ends-deliberation (Bavetta and Guala 2003, p. 435–436). This leads them to recognize the same connection between autonomy and pluralism that was recognized by Benn. If we value autonomy, we have good reason to promote a pluralistic society. Since no one person can have every sort of experience, the testimony of others who have had experiences that are very different from one's own is an invaluable means by which to increase one's own autonomy-

freedom. Only in a pluralistic society will each individual have access to such a diverse store of testimony.

According to Bavetta and Guala, then, the extent of an agent's exercise-of-autonomy freedom is simply the cardinality of the set of options he is aware of. This account is almost right; but the criterion of awareness is too stringent. The point of introducing the notion of awareness is to make room in the account of autonomy for the importance of informed choice (Bavetta and Guala 2003, p. 437). In exercising the capacity of autonomy, one makes an informed choice of a preference-ranking from among a set of possible preference-rankings. So the process of exercising the capacity of autonomy is, in part, the process of gathering relevant information about possible ends. Being free to develop this capacity is thus a matter of having access to information about possible ends.[15] The more possible ends one has access to information about, the freer one is to develop the capacity of autonomy. But one can be in possession of relevant information about a possible end without having the ability to successfully imagine what it would be like to choose that end. One need simply have evidence of one sort or another that bears on how choiceworthy the end is, as the agent in my model of ends-deliberation is represented as having. So we should identify the freedom to exercise the capacity of autonomy with the cardinality of the set of possible functionings concerning whose value the agent has some evidence.

Let us proceed, then, to examine the formal account of freedom. Both accounts have the same structure. They define $A \succeq B$ as "opportunity set A offers at least as much freedom as opportunity set B." They each propose three axioms which restrict this "as much freedom as" relation. They then each show that if and only if this relation obeys these axioms, there will be a complete, transitive and reflexive ordering of opportunity sets according to how much freedom they offer which can be represented by the cardinality of the set of *relevant* options in each opportunity set. In both cases, we can take a relevant option to be one which an agent might most prefer as a result of a well-executed course of ends-deliberation.

The first pair of axioms is:

R-M 1: $\forall x, y \in X$, if $\#\max\left(\{x\}\right) = \#\max\left(\{y\}\right)$ then $\{x\} \sim \{y\}$

S&B 1: $\forall x, y \epsilon X$ and $\forall \left(A, \Pi_i\right), \left(B, \Pi_j\right) \in P(X) \times P\left(\Pi\right)$, if $\max_i\left(A\right) = \{x\}$ and $\max_j\left(B\right) = \{y\}$ then $\left(A, \Pi_i\right) \sim \left(B, \Pi_j\right)$

In both cases, X is the universal set of options, and x and y are individual option within that set. $\{x\}$ $\{y\}$ A and B are all opportunity sets—they are all members P(X), of the power set of X. Π is the set of all deliberative preference-rankings the agent could possibly have. The members of the power set of Π (i.e. Π_i,Π_j,…) are each sets of possible preference-rankings. Π_i is the set of all the deliberative preference-rankings which are possible given some set i of options in X of which the agent is aware (or rather, concerning whose choiceworthiness the agent has some evidence; from here

[15] We must be careful in specifying what it means to "have access" to such evidence. This includes having access to the resources required to develop the basic practical rationality which one must exercise in evaluating this evidence.

on, this is what I shall mean by "being aware of an option"). $\max_i(A)$ refers to the options within the opportunity set A which an agent might most prefer as the result of a well-executed course of ends-deliberation—i.e. the options which appear at the top of at least one preference-ranking within the set Π_i. Such an option is a maximal element in A. $\#\max_i(A)$ is the number of maximal elements contained in the set A.

R-M 1 tells us that two opportunity sets $\{x\}$ and $\{y\}$ offer the same amount of effective freedom if they both contain one maximal element—i.e. an option which an agent might most prefer as the result of a well-executed course of ends-deliberation. S&B 1 tells us that the opportunity set-preference-ranking pair (A,Π_i) offers the same amount of exercise-of-autonomy freedom as the pair (B, Π_j) if A contains one maximal element according to the preference-rankings in Π_i and B contains one maximal element according to those in Π_j. The only difference between the two axioms, then, is that S&B 1 limits the agent's possible preference-rankings to those that are composed of options of which the agent is aware. It is thus somewhat more restrictive than R-M 1. But this is a welcome restriction. Let us consider the consequence of failing to adopt it.

An agent has some opportunity set which includes options which he (or some similarly situated agent) might most prefer as the result of a well-executed course of ends-deliberation, if evidence concerning the choiceworthiness of those options were available. But as it turns out, no such evidence is available for some of these options. According to Romero-Medina's account, those options still count towards the extent of the agent's effective freedom. But in this case, although the agent is free to choose these options, and although these options are valuable, the agent is not free to make an informed choice of them—the unavailability of evidence is an obstacle to his doing so. And to insist that only valuable options contribute to an agent's freedom, without requiring that the agent be free to make an informed choice among those options, is to recognize what is important about individual freedom while failing to acknowledge an important requirement for the valuable exercise of that freedom.

The second pair of axioms is:

R-M 2: $\forall A,B \in \mathrm{P}(X)$ if $A \supseteq B$ and $\max(A \setminus B) \neq 0$ then $A \succ B$; if $A \supseteq B$ and $\max(A \setminus B) = 0$ then $A \sim B$

S&B 2: $\forall (A,\Pi_i) \in \mathrm{P}(X) \times \mathrm{P}(\Pi)$ and $\forall x \in X \setminus A$ if

$\forall y \in A, \exists R_h \in \Pi_i \; s.t. \; xR_h y \;\&\; P_i s.t. \; yR_k x$ then $\left(A \cup \{x\}, \Pi_i\right) > (A,\Pi_i)$

Both of these axioms make the same claim. The extent of freedom offered by a given opportunity set increases only if a new element is added to that set which is most preferred according to some possible deliberative preference-ranking. The difference, again, is that S&B 2 requires that the agent facing the expanded opportunity set be aware of the new option.

The third pair of axioms is:

R-M 3: $\forall A,B,C,D \in \mathrm{P}(X) s.t. A \cap C = B \cap D = 0$ and $A,B,C,D \subseteq \max(X)$

if $A \succeq B$ and $C \succeq D$, then $A \cup C \succeq B \cup D$;

if $A \succeq B$ and $C \succ D$, then $A \cup C \succ B \cup D$

S&B 3:

$$\forall A,B,C,D \in \mathrm{P}(X), \forall \Pi_i, \Pi_j \in \mathrm{P}(\Pi)\, s.t. \max_i(A) \cap \max_i(C) = \max_j(B) \cap \max_j(D) = 0$$

$$\text{if } (A,\Pi_i) \geq (B,\Pi_j) \text{ and } (C,\Pi_i) \geq (D,\Pi_j), \text{ then } (A \cup C,\Pi_i) \geq (B \cup D,\Pi_j)$$

These two axioms also make the same claim. If *A*, *B*, *C*, and *D* are all opportunity sets containing maximal elements, and *A* offers at least as much freedom as *B* and *C* offers at least as much freedom as *D*, then the union of *A* and *C* offers at least as much freedom as the union of *B* and *D*. R-M 3 assumes that the opportunity sets in question are composed entirely of maximal elements, whereas S&B 3 speaks only in terms of the maximal elements within the opportunity sets, but these two formulations come to the same thing, given that it is the maximal elements in an opportunity set that count toward the freedom offered by that set on both approaches. Again, the real difference is that in the S&B model, the maximal subsets of opportunity sets are limited to options of which the agent is aware.

Both sets of axioms suffice for the same basic result. The axioms are satisfied just in case one opportunity set is said to offer at least as much freedom as another if, and only if, it contains at least as many maximal elements as the other. Satisfying either axiom set suffices for a complete, transitive, and reflexive ordering of opportunity sets according to how much freedom they offer; so we can take any two opportunity sets and compare them in this respect. We are thus assured that when we work with the notions of effective freedom and autonomy-freedom, we are able to speak precisely of the extent of freedom offered to an agent by a set of opportunities.

We have been developing a conception of negative freedom which is fit to play the role of a guiding moral and political value. Our findings thus far are as follows. Individual freedom is:

(1) The freedom to develop the capacity of autonomy. The extent of this type of freedom strictly increases with the cardinality of the set of valuable functionings concerning whose value one has access. Such access requires (a) access to evidence of the functionings' value; and (b) access to the resources required to develop basic practical rationality and self-control.

And

(2) The freedom to exercise and act on the capacity of autonomy. The extent of this type of freedom strictly increases with the cardinality of the set of those valuable functionings (a) which the agent is capable of achieving and (b) of which the agent is informed. Valuable functionings are those which might be most preferred by the agent (or a similarly situated agent) as the result of a well-executed course of ends-deliberation.

We now proceed to consider the final element of a satisfactory account of freedom.

2.5 Diversity of Choice

Martin van Hees has investigated attempts to incorporate the notion of diversity into measures of the extent of freedom offered by an opportunity set (van Hees 2004). He suggests that we conceive of how different one option in a set is from another as

a matter of how distant the two options are, as represented by a distance function. A distance function is a function d which satisfies the following four criteria:

(i) $d(x,x)=0$
(ii) $d(x,y)=d(y,x)$
(iii) $d(x,z)\leq d(x,y)+d(y,z)$
(iv) if $x\neq y$, $d(x,y)>0$

An extended distance function measures the distance between an option and a set of options. We obtain the basic extended distance function by adding a fifth criterion:

(v) $\forall x,y \in X\ \Delta(\{x\},y)=d(x,y)$

We can then develop more sophisticated metrics, according to our views on the best way to capture the distance between a set of options and a new option which might be added to the set. What we are interested in is being able to determine, given an existing opportunity set and a range of potential additions, which addition would most increase the freedom offered to an agent by the set. We could, for instance, take the distance between a new option x and a set A to be the distance between x and the member of A closest to x:

$$\Delta_{\min}(A,x)=\min_{y\in A} d(y,x);$$

or between x and the member of A farthest from x:

$$\Delta_{\max}(A,x)=\max_{y\in A} d(y,x);$$

or as a weighted average of these two:

$$\Delta^{\gamma}{}_{\mathrm{mm}}(A,x)=\gamma\left(\Delta_{\min}(A,x)\right)+(1-\gamma)\left(\Delta_{\max}(A,x)\right),0<\gamma<1.$$

And we can design still more sophisticated metrics, which take the average distance between x and the members of A:

$$\Delta_{\mathrm{av}}(A,x)=1/\#A\sum_{y\in A} d(x,y);$$

or the sum of the distances between x and the members of A:

$$\Delta_{\mathrm{sum}}(A,x)=\sum_{y\in A} d(x,y).$$

And we can combine these different approaches. Van Hees considers, for example, a combination of the min and sum approaches:

$$\Delta^{\gamma\gamma'}{}_{\mathrm{ms}}(A,x)=\gamma\left(\Delta_{\min}(A,x)\right)+\gamma'\left(\Delta_{\mathrm{sum}}(A,x)\right),\gamma>0,\gamma'>0.$$

The problem with all of these approaches, is that none of them are able to satisfy a set of plausible axioms which characterize the contribution of diversity to the

extent of freedom offered by an opportunity set. As van Hees proves, a complete, transitive and reflexive ordering of opportunity sets cannot be had from adopting any of these metrics and then assuming that the following axioms hold.

vH 1: for all distinct $x, y \epsilon X \; \{x,y\} \succ \{x\}$

Van Hees takes X to be the universal choice set. We, however, may interpret it as the set of options that would contribute to the agent's autonomy-freedom, as described above. The first axiom then simply claims that freedom increases with the addition of a relevant option.

vH 2: $\forall A, B \epsilon W$ and $\forall x \in X - \{A \cup B\}$
if $A \sim B$, then $A \cup \{x\} \succeq B \cup \{x\} \leftrightarrow \Delta(A,x) \geq \Delta(B,x)$
if $A \succ B$, then $\Delta(A, x) \geq \Delta(B,x) \rightarrow A \cup \{x\} \succ B \cup \{x\}$

The second axiom claims that if two opportunity sets off the same amount of freedom, and the same new option is added to each of them, the extent of freedom offered by the two sets will fail to remain equal if, and only if, that new option was more distant from one set than it was from the other. If one set offers more freedom than another, it will continue to do so as long as the new option is at least as distant from it as from the other set.

vH 3: $\forall A, B \epsilon W$ and $\forall x \notin A$ and $\forall y \notin B$
if $A \sim B$, then $A \cup \{x\} \succeq B \cup \{y\} \leftrightarrow \Delta(A,x) \geq \Delta(B,y)$
if $A \succ B$, then $\Delta(A, x) \geq \Delta(B, y) \rightarrow A \cup \{x\} \succ B \cup \{y\}$

The third axiom claims that if two opportunity sets off the same amount of freedom, and new and *different* options are added to each one, then the extent of freedom offered by the two sets will fail to remain equal if, and only if, there is a discrepancy in the distance between the sets and their new options. If one set offers more freedom than another, it will continue to do so as long as the option added to it is not closer to it than the option added to the other set is to the other set.

vH 4: There are positive numbers $g, k \; (k > g) s.t. \forall A \epsilon W$ and $\forall x, y \notin A$
if $\Delta(y,z) \geq k \forall z \in A$, and $\Delta(x,z) \leq g$ for some $x \in A$, then $A \cup \{y\} \succeq A \cup \{x\}$

The fourth axiom is what van Hees calls the principle of insensitivity to small differences. At some point, a new option is so close to an option already included in the opportunity set, that it cannot increase the freedom offered by that set more than another new option which is not that close to any already included option.

As noted above, none of the metrics introduced by van Hees can satisfy all of these axioms. Since all the axioms are fairly plausible, this is bad news for the suggested metrics. And without a metric capable of satisfying plausible axioms to yield a proper ordering of opportunity sets, we have no hope of integrating diversity into our conception of individual freedom. So we need a new metric. In order to begin constructing one, we need a clear idea of the source of the trouble for the metrics already introduced. It will be most instructive to examine what goes wrong for Δ_{av} and Δ_{sum}. Suppose we have a set A with three equally spaced options:

$$x \quad y \quad z$$

Then we add a new option w which is very close to x (i.e. $d(w,x) \leq g$):

$$wx \quad y \quad z$$

Consider the two sets $\{x,y,z\}$ and $\{w,x,z\}$. If we use either Δ_{av} or Δ_{sum}, then according to vH 3, $\{w,x,z\}$ offers more freedom than $\{x,y,z\}$. But according to vH 4, $\{x,y,z\}$ offers more freedom than $\{w,x,z\}$.

The problem with the distance metrics introduced thus far is that they are all borrowed from other contexts in which the notion of distance receives a very different interpretation from the one that is relevant for our purposes. By "distance," we mean how different one option is from another, and we are interested in finding a way to capture the extent to which adding a new option will increase the diversity of the options offered by an opportunity set. So we need to start by sharpening our understanding of the notion of diversity. First, suppose a new option is introduced which lies somewhere between two prior options:

$$x \quad y \quad z$$

The position of y which adds the most diversity in this case is the position which is exactly in between x and z. If there are another two options a and b which are even farther apart than x and z, then placing y exactly in between them will introduce even more diversity:

$$a \quad y \quad b \quad x \quad z$$

And it seems plausible to assume that y need not be exactly in between a and b for its location between them to be preferred to a location exactly in between x and z. It need only be the case that y is at least as far from either a or b then it would be from either x or z, were it located exactly between x and z.

But of course, the new option need not be located in between any two prior options. It might extend the "length" (so to speak) of the opportunity set itself:

$$y \quad a \quad b \quad x \quad z$$

The main problem we face in constructing a metric adequate to measure diversity of choice is that there does not seem to be any single formula capable of measuring the contribution to diversity of a new option in both of these situations. In these last two cases, if we assume that the distance between y and a is $> g$ in both, then Δ_{av} and Δ_{sum} will consider the second set to be more diverse than the first. But if y is still relatively close to a in the second case, and quite far from a and b in the second, we may be justified in resisting this result. We need to acknowledge that there is an important difference between the way diversity is enhanced by the addition of options internal to a set and by the addition of options which extend a set.

Let us define the "neighbors" of a potential new option to be the prior options which would be located directly to its left and right. More precisely, given a set A and a potential new option x_n:

$$x_n \text{is a neighbor of } x_0 \epsilon A =_{\text{def.}} -\exists y \in A s.t. d\left(x_0, y\right) + d\left(y, x_n\right) = d\left(x_0, x_n\right)$$

New options which are internal to a set will have two neighbors, whereas options which extend a set will have one. We then define a new distance metric:

$$\Delta_{dc}\left(x, A\right) = \{^{\Delta^1_{dc}}_{\Delta^2_{dc}}$$

We use $\Delta^1_{dc}(x,A)$ to measure the distance between a new option x and a set A when x has one neighbor in A, and $\Delta^2_{dc}(x,A)$ when it has two. We then define $\Delta^1_{dc}\left(x_0, A\right)$ as:

$$\Delta \min\left(x_0, A\right) \cdot \frac{\Delta_{\max}\left(x_0, A\right)}{\Delta_{\max}\left(x_0, A\right) - \Delta_{\min}\left(x_0, A\right)}$$

On the left hand side of this expression, we have the distance between the new option and the nearest prior option. On the right, we have the ratio of the distance between the new option and the farthest prior option, and the length of the string of prior options. This metric represents a plausible approach to measuring the extent to which the diversity of an opportunity set has been extended. We take the increase in the length of the string which results from adding the new option, and multiply it by a ratio which expresses the factor by which the length of the string is increased due to the addition of the new option. Suppose $A = \{x,y\}$ and $d(x,y) = 2$. We then add z, such that $d(z,x) = 1$ and $d(z,y) = 3$:

$$z \quad x \quad y$$

The addition of z lengthens the string by 1, and increases the length of the prior string by a factor of 3/2. The metric thus takes into account both the novelty of z (how far it is from anything already offered by the opportunity set—in this case, 1 unit of distance), and the extent to which the length of the whole string is increased by the addition of z.

For internal additions, we define $\Delta^2_{dc}\left(x_0, A\right)$ as:

$$\frac{d\left(x_0, x_n^1\right) + d\left(x_0, x_n^2\right)}{2} \cdot \frac{d\left(x_0, x_n^1\right)}{d\left(x_0, x_n^2\right)} suchthat \frac{d\left(x_0, x_n^1\right)}{d\left(x_0, x_n^2\right)} \leq 1$$

The left-hand side of this expression is the average distance between the new option and its two neighbors. The right-hand side is the ratio of these two distances, with the larger distance in the denominator. The metric thus takes two factors into account. First, greater diversity is added when we add an option in between prior

option which are far apart. Second, greater diversity is added when the new option is close to being exactly in between the two prior options. The left-hand ratio takes the first of these factors into account, and the right-hand ratio takes the second.

I have not worked out whether this distance metric manages to yield a complete, transitive and reflexive ordering given van Hees' axioms. But it does seem to resolve the problem encountered by Δ_{av} and Δ_{sum}. Again, consider the string

$$w \; x \quad y \quad z$$

and the sets $\{w,x,z\}$ and $\{x,y,z\}$. Assume $d(x,y) = d(y,z) = 1$, and $d(w,z)$ is some very small number $\leq g = 0.1$. Then $\Delta_{dc}(w,\{x,z\}) = (.1)(2.1/2) = 0.105$, and $\Delta_{dc}(y,\{x,z\}) = (1)(1) = 1$. And $\{x,y,z\}$ will offer more freedom than $\{w,x,z\}$ according to both vH 3 and vH 4 (which is the result we wanted but failed to obtain for Δ_{av} and Δ_{sum}).

3 Reconciling Autonomy-Freedom and Diversity

If my distance metric does succeed in capturing the notion of diversity of choice within an opportunity set, then only one task remains to complete the development of the conception of negative freedom we have been searching for. Autonomy-freedom has been characterized by a set of axioms governing the "offers as much freedom as" relation. Another set of axioms has been proposed by van Hees which govern this relation with respect to the impact of introducing greater diversity. But these two sets of axioms are inconsistent. In particular, S&B 3 conflicts with vH 2–4. To see this, consider the following case. A and B are both maximal sets, $A \cap B = 0$, and $A \succeq B$. $C = \{x\} \notin A, B$ and $D = \{y\} \notin A, B$. C and D are both maximal sets. According to S&B 3, it should therefore be the case that $A \cup C \succeq B \cup D$. But suppose $\Delta_{dc}(x,A)$ is very small and $\Delta_{dc}(y,B)$ is very large. We might then want to claim that $B \cup D \succ A \cup C$.[16] The final task is to reconcile these two measures of freedom.

A promising beginning has been made by Antonio Romero-Medina and Vito Peragine (Romero-Medina and Peragine 2006). Their strategy is to define a reflexive and symmetric binary relation S over the choice set X: xSy is to be interpreted as "option x is similar to option y." A choice set $A \in P(X)$ will then be said to be a *homogenous* choice set iff $\forall a, a' \in A, aSa'$. A *similarity-based partition* of a choice set A can then be formed by breaking the set A up into subsets such that each subset is homogenous. There may be many similarity-based partitions of a single choice set, and these are denoted as $\phi(A)$ $\phi'(A)$ $\phi''(A)$, etc. $\Phi(A)$ is the set of all similarity-based partitions $\phi(A)$ such that for every similarity-based partition $\phi'(A)$, $\#\phi'(A) \geq \#\phi(A)$. So $\Phi(A)$ is the set of all the smallest similarity-based partitions of A—it is the

[16] Every approach to diversity of choice must assume that there is an objective fact regarding how similar or different one available option is to another. This task seems to me to be fairly manageable for opportunity sets consisting of functionings, at least when those functionings are described in a suitably general way.

set of all the ways of breaking A up into homogenous subsets, such that the fewest number of subsets possible are needed. The subsets in these partitions will thus contain more elements than are contained in the subsets found in any partitions that do not belong to this set.

An option x will be said to be similar to a choice set A iff $xSa \forall a \in A$. Finally, for all $A, B \in P(X)$ with A homogenous, A does not *mimic* B iff, for all $\phi(B)$ $\Phi(B)$, there exists $a \in A$ such that, for all subsets $B_i \in \phi(B)$, it is not the case that aSB_i. In other words, a homogenous choice set A fails to mimic a choice set B iff there is at least one element in A which is not similar to any subset of any smallest similarity-based partition of B.

This terminology is introduced so that the following four axioms may be stated (van Hees 2004, pp. 33–34):

R-M & P 1: $\forall x, y \in X, \forall \Pi_i, \Pi_j \in P(\Pi), (x, \Pi_i) \sim (y, \Pi_j)$

This first axiom states that two sets of one single option each always offer an equal degree of freedom.

R-M & P 2:

$$\forall A \in P(X), \forall \Pi_i \in P(\Pi), \forall x \in X \text{ such that } \max_i(A \cup \{x\}) = \max_i(A), (A \cup \{x\}, \Pi_i) \sim (A, \Pi_i)$$

The second axiom simply states that adding an option which is not a maximal option according to any reasonable preference profile does not increase the amount of freedom offered by a choice set.

R-M & P 3: $\forall A \in P(X), A \text{ homogenous}, \forall \Pi_i \in P(\Pi), \forall x \in X - A \text{ such that } x \in \max_i(A \cup \{x\})$,

$$[xS \max_i(A)] \rightarrow (A \cup \{x\}, \Pi_i) \sim (A, \Pi_i)$$

And

$$[x - S \max_i(A)] \rightarrow (A \cup \{x\}, \Pi_i) > (A, \Pi_i)$$

So according to the third axiom, if a choice set A is homogenous, then adding an option x which is similar to A will not increase the freedom offered by A, even if x is a maximal element according to some reasonable preference profile. But if x is not similar to A, then adding it will increase the freedom offered by A if it is a maximal element.

R-M & P 4:

$$\forall A, B, C, D \in P(X), \forall \Pi_i, \Pi_j \in P(\Pi), \text{such that } C \text{ and } D \text{ are homogenous and } \max_i(C \cup A)$$
$$\text{does not mimic } \max_i(A), \text{and } \max_i(A) \cap \max_i(C) = \max_j(B) \cap \max_j(D) = \varnothing,$$
$$[(A, \Pi_i) \geq (B, \Pi_j) \text{and} (C, \Pi_i) \geq (D, \Pi_j)] \rightarrow [(A \cup C, \Pi_i) \geq (B \cup D, \Pi_j)]$$
$$[(A, \Pi_i) > (B, \Pi_j) \text{and} (C, \Pi_i) \geq (D, \Pi_j)] \rightarrow [(A \cup C, \Pi_i) > (B \cup D, \Pi_j)]$$

This final axiom states that if one choice set offers at least as much freedom as another, it will continue to do so when a homogenous, non-mimicking choice set is added to each of the original sets, so long as the set added to the first original set offers as much freedom as the set added to the second original set. If one choice set offers more freedom than another, it will continue to offer more freedom under this condition.

Romero-Medina and Peragine prove that if a ranking of opportunity sets obeys these four axioms, then the following holds (van Hees 2004, p. 34):

$$\left(A, \Pi_i\right) \geq \left(B, \Pi_j\right) \leftrightarrow \#\varphi\left(\max_i\left(A\right)\right) \geq \#\varphi\left(\max_j\left(B\right)\right)$$

In order to rank any two opportunity sets according to how much freedom they offer, then, we look at the smallest similarity-based partition of the set of maximal elements from each of those two sets. We then look at how many subsets are contained within each of those smallest similarity-based partitions. The set whose partition contains a greater number of subsets offers more freedom.

I have described this approach to reconciling the cardinality-approach and the diversity-approach to measuring freedom as a promising beginning. The main difficulty that plagues it is that there is no reason to interpret the binary relation S employed by Romero-Medina and Peragine as a relation of similarity between options in any meaningful sense, given that they only require that it be reflexive and symmetric. Much more than this is required, if we are to be justified in interpreting this relation as a similarity relation in any relevant sense. Fortunately, we have somewhere to turn to for guidance in articulating the conditions which a binary similarity relation should satisfy, guidance that was lacking in the case of the similarity function. Ariel Rubinstein has offered a set of six requirements that do seem to adequately characterize a binary similarity relation (Rubinstein 1998, p. 29). Define the similarity relation, ~, as a binary relation on the set $I=[0,1]$. The restrictions Rubinstein proposes are:

Reflexivity: for all $a \in I, a \sim a$.
Symmetry: for all $a, b \in I$, if $a \sim b$ then $b \sim a$.
Continuity: the graph of ~ is closed in $I \times I$.
Betweenness: if $a \leq b \leq c \leq d$ and $a \sim d$, then $b \sim c$.
Nondegeneracy: $-\left(0 \sim 1\right)$ and for all $0 < a < 1$ there are b and c so that $b < a < c$ and $a \sim b$ and $a \sim c$. For $a = 1$ there is $b < a$ so that $a \sim b$.
Responsiveness: Denote by a^* and a_* the largest and smallest elements in the set that are similar to a. Then a^* and a_* are strictly increasing functions (in a) at any point at which they obtain a value different from 0 or 1.

This is the point at which the above discussion of distance metrics and the formulation of a metric designed specifically to capture our intuitions about diversity of choice pays off. If we interpret "*a* is similar to *b*," as meaning "the distance between *a* and *b*, *d(a,b)*, is less than some maximum distance *m*," we will have an interpretation of the similarity relation that fulfills all of Rubinstein's requirements.

The similarity-based partitions that figure in Romero-Medina's and Peragine's model, therefore, should be taken to be subsets of a choice set, none of whose elements exceed the maximum distance from one another. And we may extend this idea to include similarity between a new option and a choice set: $xSA \leftrightarrow \Delta(x,A) \leq \mu$. We may then take the extended distance metric Δ to be my metric Δ_{dc}. If we understand the notion of similarity between a new option and a prior choice set in this way, we will both be justified in taking ourselves to have successfully captured a notion of similarity which is relevant in this context (since all of Rubinstein's requirements will be satisfied), and to have isolated a metric for measuring the distance between options and sets which shows maximum respect to our intuitions about the way in which freedom is increased by the addition of greater diversity of choice.

My final proposal, then, is that we replace Romero-Medina's and Peragine's notion of similarity between an option and a set, *xSA*, with a limit on the distance between an option and a set as measured by my distance metric, $\Delta_{dc}(x,A) \leq \mu$.[84] We can then avail ourselves of their axioms and the decision rule that follows from them. Their decision rule suffices for a complete, transitive and reflexive ordering of opportunity sets according to how much freedom they offer. My account of ends-deliberation allows us to specify exactly what it means for an option to count as maximal according to a reasonable preference profile, and my distance metric provides a precise and plausible way of cashing out the idea that an agent's freedom never increases with the addition of an option which is too much like other options already available to the agent. The result is a single model for measuring the amount of freedom offered by opportunity sets which is both wide-ranging in the considerations it takes into account, and philosophically satisfying in the way it incorporates those considerations.

4 The Freedom to Exercise One's Autonomy: A Two-stage Approach

Our search for a conception of individual negative freedom which is fit to play the role of a guiding moral and political value is nearing an end. The conception of individual freedom that has been developed thus far may be summarized as follows:

(1′) The freedom to develop the capacity of autonomy. The extent of this type of freedom strictly increases with the cardinality of the set of valuable functionings concerning whose value one has access. Such access requires (a) access to evidence of the functionings' value; and (b) access to the resources required to develop basic practical rationality and self-control.

And

(2′) The freedom to exercise and act on the capacity of autonomy. The extent of this type of freedom strictly increases with the cardinality of the set of those valuable functionings (a) which the agent is capable of achieving; (b) of which the

agent is informed; and (c) which contribute sufficiently to the diversity of the agent's capability set. Valuable functionings are those which might be most preferred by the agent (or a similarly situated agent) as the result of a well-executed course of ends-deliberation.

There is, however, an important deficiency in (2′), which purports to define the extent of an agent's freedom to exercise his autonomy. This freedom is important at two stages in the life of the agent who has developed the capacity of autonomy. The first stage is in the choice of which set of capabilities to develop, a choice made from a set of capability sets. Each functioning in each accessible capability set will be one which the agent has the potential to achieve; and each accessible capability set will consist of functionings which the agent has the potential to achieve jointly. But a choice of capability set—a choice of which capabilities to develop together, and which to neglect—is necessitated by the fact that the agent cannot jointly achieve every functioning which he has the potential to achieve individually. The second stage is in the choice of which functionings to strive to achieve from within the particular set of capabilities the agent has developed. But only the extent of the agent's freedom in making the latter choice is defined in (2′).

This deficiency, however, is (mostly) easy to remedy. We can define the extent of the agent's freedom to exercise his autonomy at the first stage as the cardinality of the set of the accessible, sufficiently large, and sufficiently diverse capability sets of which he is informed. An accessible capability set, Q_a, is one the agent has the natural potential to develop. Call such a set sufficiently large if the number of its (sufficiently diverse) members exceeds some threshold, so that excessively restrictive capability sets do not end up counting toward the measure of the agent's freedom of choice among potential capability sets. Call an accessible, sufficiently large, and sufficiently diverse capability set a maximal capability set, for an agent i, if it is a capability set which that agent, or a similarly situated agent, might most prefer to develop after a well-executed course of ends-deliberation. Call the set of all maximal capability sets $\mathcal{Q}$. So we have: $\forall aQ_a \in \mathcal{Q} iff\ \exists i such that \#\phi\left(max_i\left(Q_a\right)\right) > n$. A set of maximal capability sets, $\boldsymbol{Q}_a$, is a subset of $\mathcal{Q}: Q_a \subseteq \mathcal{Q}$

Deliberation over potential capability sets proceeds in a straightforward way. Suppose the agent has successfully developed the capacity of autonomy—having had the freedom to do so per (1′)—and has, at the time in his life $t = 1$, arrived at a first deliberative preference-ranking $R^*_{t=1}$, in accordance with (ED*).[17] Each functioning appearing in $R^*_{t=1}$ will have a valuation based on its position in the ranking.

[17] As the agent's life proceeds, he may of course alter his views regarding the expected values of various ends. Such alterations may take place after the agent has decided what capabilities to develop, after he has developed them, and after he has achieved some or all of the functionings of which he is capable. He may end up deciding to neglect some of his capabilities, and develop and exercise new ones later in life. In particular, though, the agent will be expected to engage in another round of ends-deliberation about which functionings to strive to achieve, once he has developed a particular set of capabilities—the freedom to do which is defined in (2′). We can refer to this later deliberative preference-ranking as $R^*_{t=2}$.

We need only define the value, for the agent, of a particular capability set as the sum of the valuations of the achieved functionings contained in that set:

$$v\left(Q_a\right) = \Sigma_{f \in Q_a} v\left(f\right).$$

We will then have the following principle of choice over capability sets:

$$\left(\mathrm{CC}\right): Q_a \geq_i Q_b \leftrightarrow v_i\left(Q_a\right) \geq v_i\left(Q_b\right), \textit{for an agent } i.$$

We are concerned, then, with the sets of maximal capability sets of which an agent is informed: $\forall a, b \forall i, j \left(Q_a, \Pi_i\right)\left(Q_b, \Pi_j\right)$. It is straightforward to adapt Bavetta and Guala's axioms to measure the extent of freedom of choice among capability sets offered by a set of capability sets, and to compare the extent of freedom of choice offered by any two such sets. The extent of an agent's freedom to exercise his autonomy by choosing among potential capability sets from a set of such sets will indeed then strictly increase with the cardinality of the set of capability sets:

$$\left(Q_a, \Pi_i\right) \geq \left(Q_b, \Pi_j\right) \leftrightarrow \#\left(\max_i\left(Q_a\right)\right) \geq \#\left(\max_j\left(Q_b\right)\right).$$

One problem does of course remain. We would ideally like some way to measure the diversity of the set of capability sets—some way to measure how different the accessible capability sets are from one another—so that we don't end up counting both a two accessible capability sets which are nearly, but not exactly, the same. That is, we would like to be able to define $\#\phi(max_i(\boldsymbol{Q}_a))$. This would require a proposal for a distance metric between multi-member sets which could plausibly be interpreted as measuring how different those multi-member sets are from one another. But I know of no such proposal, nor, unfortunately, do I have a plausible one of my own to offer. So this is a weakness in the account which we shall have to accept, at least at present.

Our definition of individual freedom can now take on its final form:

(1*) The freedom to develop the capacity of autonomy. The extent of this type of freedom strictly increases with the cardinality of the set of valuable functionings concerning whose value one has access. Such access requires (a) access to evidence of the functionings' value; and (b) access to the resources required to develop basic practical rationality and self-control.

(2*) The freedom to exercise and act on the capacity of autonomy (stage one). The extent of this type of freedom strictly increases with the cardinality of the set of capability sets which are themselves sufficiently large and diverse and accessible to the agent. A capability set is accessible to an agent when (a) the agent has the natural potential, and the resources, needed to jointly develop the capabilities in that set; and (b) the agent is informed about the capabilities in that set.

(3*) The freedom to exercise and act on the capacity of autonomy (stage two). The extent of this type of freedom strictly increases with the cardinality of the set of those valuable functionings (a) which the agent is capable of jointly achieving; (b) of which the agent is informed; and (c) which contribute sufficiently to the diversity of the agent's capability set. Valuable functionings are those which might be most preferred by the agent (or a similarly situated agent) as the result of a well-executed course of ends-deliberation.

5 Conclusion: A Compound Conception of Liberty

I have now developed rigorous and precise accounts of the capacity of autonomy and of the freedom to develop, exercise and act on that capacity. This finally makes possible a definition of a life of liberty, which brings these notions together:

> An individual agent leads a life of liberty when he (a) has the freedom to develop his capacity of autonomy, as defined in (1*); (b) succeeds in developing that capacity; (c) has the freedom to exercise and act on his autonomy by choosing what capabilities to develop and working to develop them, as defined in (2*); (d) succeeds in exercising and acting on that capacity by deliberating well over his set of accessible capability sets and choosing one to develop, in accordance with (ED*) and (CC), by choosing how to cultivate that capability set, *via* instrumental reasoning, and by exercising self-control to maintain his prior intention to pursue that capability set when there is no reason to revise it and to act on that intention when the time comes; (e) has the freedom to exercise and act on his autonomy by choosing what functionings within his capability set he will pursue and pursuing them, as defined in (3*); and (f) succeeds in exercising and acting on that capacity by deliberating well over his set of accessible functionings and choosing which to pursue, in accordance with (ED*), by choosing how to pursue those functionings, *via* instrumental reasoning, and by exercising self-control to maintain his prior intentions to pursue those functionings when there is no reason to revise them and to act on those intentions when the time comes.

We can interpret the life of liberty from an Aristotelian perspective as a progression from potentialities to actualities. The agent begins with the natural potential to develop the capacity of autonomy in its various aspects—to develop basic practical rationality and the capacity for self-control. This, in Aristotelian terminology, is a first potentiality. The agent likewise begins with a first potentiality to develop a host of other abilities. If the agent has access to resources, and to evidence regarding the value of potential ends and the effectiveness of potential means, which are required to develop the capacity of autonomy in addition to this first potentiality, we say the agent has the capability-freedom to develop the capacity of autonomy. This is the freedom described in (1*) above. The agent who successfully develops the capacity of autonomy moves from a first potentiality to a second potentiality/first actuality. Likewise, the agent who has the resources needed to develop his other natural potentials, and to exercise the ones he chooses to develop, has the capability-freedom to develop those abilities. The agent who successfully develops some set of those abilities moves from a first potentiality to a second potentiality/first actuality. The agent who has already developed the capacity of autonomy, and who has evi-

dence regarding the value of the abilities he could choose to develop, has the capability-freedom to autonomously choose which of those other abilities to develop, and to act autonomously on that choice. This is the freedom described in (2*). If he does so, he thereby exercises his autonomy—a move from a first to a second actuality. The agent who has developed a set of abilities and has the resources to exercise them has the capability-freedom to achieve those functionings. The agent who exercises his abilities moves from a first to a second actuality. The agent who has developed the capacity of autonomy, and has evidence regarding the value of the functionings that can be achieved by exercising them, has the capability-freedom to autonomously choose which of these abilities to exercise and act autonomously on that choice. This is the freedom described in (3*).

One of the greatest virtues of this account of liberty is that it is measurable. We can directly compare how much freedom, of any of the types described above, any two individuals possess. And given a precise decision-theoretic account of autonomy like mine, we can also compare the level of autonomy achieved by any two individuals, though much works remains to be done to develop a suitable operational metric. There are two strategies for doing this. The first emphasizes autonomy as a (set of) functioning(s), and seeks to measure the level at which an individual is achieving functionings (Stewart 1991; Brandolino and D'Alessio 1998; Kuklys 2005, ch. 2–3; Wolff and De-Shalit 2013, ch. 6). The second emphasizes autonomy as an exercise of rationality, and seeks to measure an individual's deviation from the ideal in that exercise (Jones 1999). Whichever type of metric we judge most promising, denoting an individual's level of autonomy as A, and his freedom to develop his autonomy, choose autonomously among capability sets to develop, and choose autonomously to exercise his capabilities, as Q_A, Q_D, and Q_E respectively, we can express the extent of any agent's liberty as:[18]

$$L = Q_A + A\left(Q_D + Q_E\right).$$

Given a measurable conception of liberty, we can speak precisely in terms of the distribution of liberty within a society. My next task is to argue that the distribution of liberty, in precisely the way that I have conceived of it, ought to be the central concern of a liberal society.

[18] This formula is offered as an illustration of the fact that we are now working with a conception of liberty which is, at least theoretically, measurable. Because the theory of distributive justice I will go on to develop is not an aggregate maximizing view of any kind—not even one that proposes maximizing the aggregate of individual liberty—the adequacy of this illustrative formulation is of limited significance. My view focuses instead on comparisons between individual's achievement of each element of a life of liberty, defeasibly prioritized. This requires that each element of the life of liberty be measurable, but does not assign any theoretical role to comparisons between individual's overall liberty.

References

Aristotle. *Nicomachean ethics* [Ethica Nicomachea] Oxford classical texts, ed. L. Bywater. Oxford: Oxford University Press.

Arneson, R. 1985. Freedom and desire. *Canadian Journal of Philosophy* 15(3): 425–448.

Arpaly, N. 2000. On acting rationally against one's better judgment. *Ethics* 110: 488–513.

Barnes, J. (ed.). 1984. *The complete works of Aristotle: The revised Oxford translation, Vol. I.* ed. Jonathan Barnes. Princeton: Princeton University Press.

Baumeister, R.F. 2002. Ego depletion and self-control failure: An energy model of the self's executive function. *Self and Identity* 1(2): 129–136.

Baumeister, R.F., and K.D. Vohs. 2007. Self-regulation, ego depletion, and motivation. *Social and Personality Psychology Compass* 1: 115–128.

Baumeister, R.F., E. Bratslavsky, M. Muraven, and D.M. Tice. 1998. Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74: 1252–1265.

Baumeister, R.F., C.N. Dewall, N.J. Ciarocco, and J.M. Twenge. 2005. Social exclusion impairs self-regulation. *Journal of Personality and Social Psychology* 88: 589–604.

Baumeister, R.F., et al. 2006. Self-regulation and personality: How interventions increased regulatory success, and how depletion moderates the effects of traits on behavior. *Journal of Personality* 746: 1773–1801.

Baumeister, R.F., E.A. Sparks, T.F. Stillman, and K.D. Vohs. 2008. Free will in consumer behavior: Self-control, ego depletion, and choice. *Journal of Consumer Psychology* 18: 4–13.

Bavetta, S., and F. Guala. 2003. Autonomy-freedom and deliberation. *Journal of Theoretical Politics* 15(4): 423–443.

Berlin, I. 1969/2006. Two concepts of liberty. In *Contemporary political philosophy: An anthology*, ed. R.E. Goodin, and P. Pettit, 369–386. New York: Blackwell.

Brandolino, A., and G.D'Alessio. 1998. Measuring well-being in the functionings space. Working Paper Series, Banca D'Italia.

Bratman, M. 1979. Practical reasoning and weakness of the will. *Nous* 13: 153–171.

Davidson, D. 1980. How is weakness of the will possible? In *Essays on actions and events*, 21–42. Oxford: Clarendon Press.

Heath, J. 2008. *Following the rules: Practical reasoning and deontic constraint*. Oxford: Oxford University Press.

Holton, R. 1999. Intention and weakness of will. *Journal of Philosophy* 96: 241–262.

Hurka, T. 1993. *Perfectionism*. Oxford: Oxford University Press.

Jones, B.D. 1999. Bounded rationality. *American Review of Political Science* 2: 297–321.

Kuklys, W. 2005. *AmartySen, A's capability approach: Theoretical insights and empirical applications*. Berlin: Springer.

MacIntyre, A. 2006. What is wrong with weakness of will? *Journal of Philosophy* 103: 284–311.

Mele, A. 1987. *Irrationality*. New York: Oxford University Press.

Moldoveanu, M. 2011. *Inside man: The discipline of modeling human ways of being*. Palo Alto: Stanford Business Press.

Pattanaik, S., and Y. Xu. 1998. Preference and freedom. *Theory and Decision* 44(2): 173–198.

Pettit, P. 2001. Capability and freedom: A defense of Sen. *Economics and Philosophy* 17(1): 1–20.

Romero-Medina, A. 2001. More on preferences and freedom. *Social Choice and Welfare* 18(1): 179–191.

Romero-Medina, A., and V. Peragine. 2006. On preference, diversity and freedom. *Social Choice and Welfare* 27(1): 29–40.

Rubinstein, A. 1998. *Modeling bounded rationality*. Cambridge, MA: MIT Press.

Sen, A. 1999. *Commodities and capabilities*. Oxford: Oxford University Press.

Sen, A. 2002. *Rationality and freedom*. Cambridge, MA: Harvard University Press.

Stewart, A.L. (ed.). 1991. *Measuring functioning and well-being*. Durham: Duke University Press.

Tenenbaum, S. 1999. The judgment of a weak will. *Philosophy and Phenomenological Research* 59: 875–911.

van Hees, M. 2004. Freedom of choice and diversity of options: Some difficulties. *Social Choice and Welfare* 22(1): 253–266.

Vohs, K.D., R.F. Baumeister, and B.J. Schmeichel. 2012. Motivation, personal beliefs, and limited resources all contribute to self-control. *Journal of Experimental Social Psychology* 48: 943–947.

Wolff, J., and A. De-Shalit. 2013. *Disadvantage*. Oxford: Oxford University Press.

Part III
Justice: Distribution

Chapter 8
Justice: Distribution – Introduction

Over the course of the next several chapters, I will criticize a number of theories of distributive justice, and introduce and defend my own. The theory I will argue for, which I call the Theory of Equal Liberty, is a variety of egalitarianism. In essence, it claims that the appropriate distributive role of the State is to guarantee that each citizen enjoys the greatest possible equal share of liberty, as this notion has been defined over the course of the last four chapters. This is why it is so important that our concept of liberty be measurable—the core claim of the Equal Liberty approach to distributive justice would be meaningless if it were not possible to determine whether, for any element of liberty and for any two individuals, one enjoyed a greater or lesser degree of that element of liberty than the other, or whether they were equal in this respect.

There are many ways of approaching the issue of distributive justice, and for each type of approach, there are many individual theories. The aim of this book is primarily a positive, rather than a critical one—my goal is to articulate and defend a particular conception of liberalism, not to argue definitively against every other kind. In this introductory chapter, therefore, I will try to justify the rather severely limited range of theories which I will discuss and criticize in depth in the next chapter. The way to do this is first to ascend to a fairly high level of abstraction, and provide a taxonomy of the types of theories of distributive justice. I will then give some very general reasons for finding entire types of theories unsatisfactory—which are by no means meant to be taken as decisive arguments against the theories of those types—and thus, I hope, justify my decision not to subject a number of prominent theories to more thorough examination and criticism in what follows.

The most general division of theories of distributive justice is between procedural theories and goal-based theories. This distinction does not track Robert Nozick's distinction between historical theories and pattern-based theories. I assume that vanishingly few individuals, if indeed any at all, are still captivated by Nozick's hopeless "Wilt Chamberlain" argument.[1] Any theory of distributive justice

[1] For the argument, see (Nozick 1974, pp. 160–163). For a thorough and enlightening critique, see (Cohen 1995, ch. 1).

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_8

worth taking seriously recognizes that the structure of legal, political, and social institutions must be specified at the outset, and then allowed to function as designed and within the established parameters. Goal-based theories are those which justify institutional design by appealing to a social goal—which itself must be justified—to which such institutions are conducive. Utilitarianism, considered as a theory of distributive justice, is certainly a goal-directed theory; but no serious utilitarian advocates the kind of constant interference with individual choices that Nozick, rather inexplicably, saw as part and parcel of any goal-based theory. The utilitarian, when he turns his attention to distributive justice, is interested in what sort of legal, political and social institutions, the design of which he seeks to specify at the outset, are most likely to maximize aggregate utility across the society. One of these institutions is of course the tax code, and a utilitarian tax code will be designed to redistribute income with the utilitarian's goal in mind. The specifics of the tax code to which one is subject will of course have some impact on one's incentives; but we are nowhere near the neighborhood of the continual direct interference with individual decisions about individual purchases, or with the freedom of individuals to enter into particular mutually agreeable contracts, that Nozick frets over. There is a legitimate concern about whether it is morally permissible for the State to shape individual incentives through policy. I address this concern in Chap. 16, where I argue that there is nothing morally impermissible about the State pursuing the perfectionist policies required by Equal Liberty.

1 Contract Theories

Procedural theories of distributive justice are social contract theories. The basic idea behind a social contract theory is that what makes a just society just is not that it satisfies some independently defensible criterion of social justice, but that it is the kind of society that individuals, in the right setting, would agree to form and live in. Social contract theories can be divided into ideal contract theories (often called "contractualist" theories) and non-ideal contract theories (often called "contractarian" theories). For those who are fairly well steeped in the distributive justice literature, the easiest way to illustrate these distinctions is by identifying each type of theory with its most prominent expositors. Contractarian theories come in two main flavors: Hobbesian (best illustrated by David Gauthier) (Gauthier 1986); and Lockean (best illustrated by Nozick) (Nozick 1974). Contractualist theories are subdivided into veil-of-ignorance theories and ideal-speech-situation theories. The most prominent advocates of ideal-speech-situation contractualism are Brian Barry, T.M. Scanlon, Gerald Gaus, and Jürgen Habermas. And finally, veil-of-ignorance contractualism comes in social-choice-theoretic and Kantian variants. John Harsanyi's purported argument for utilitarianism is a paradigmatic example of a

choice-theoretic veil-of-ignorance theory.[2] Rawls' theory of Justice as Fairness was originally presented as *both* choice-theoretic and Kantian (Rawls 1971, ch. 3–4). Rawls seems to have originally conceived of his theory as essentially choice-theoretic, but as compatible with a Kantian interpretation. Before long, however, it became clear that Rawls had not in fact developed a choice-theoretic contractualism at all.[3] In his subsequent work, Rawls moved more and more in the direction of Kantian contractualism (Rawls 1987, 2001).

The essential difference between contractarian and contractualist theories is that contractarian theories only assume that the agents who come together to form the social contract are rational, self-regarding, and free from coercion and manipulation. Contractualist theories put more stringent requirements on the structure of the setting in which the social contract is hashed out and assume more of the agents involved (and differ based on differences in these requirements). Hobbesian contractarian theories assume that the agents who enter into the social contract are "rational," in the sense in which this term is used in neoclassical economics: they are *homines economici*, individuals with exogenously given, complete, transitive, self-regarding preference-rankings, represented by concave utility functions, who agree to enter into a social contract with an eye only to maximizing their own individual expected utility. There are a number of reasons why no Hobbesian contractarian theory can be an adequate theory of distributive justice. My primary reason for making this sweeping claim has to do with what I take to be the first of the fundamental criteria which any adequate theory of distributive justice must satisfy: avoiding the error which lies at the heart of all forms of what Frank H. Knight calls "liberal individualism":

> These reflections naturally lead up to the most important single defect, amounting to a fallacy, in liberal individualism as a social philosophy. The most general and essential fact that makes such a position untenable is that *liberalism takes the individual as given*, and views the social problem as one of right relations between given individuals. This is its fundamental error. The assumption that this can be done runs counter to clear and unalterable facts of life. The individual cannot be a datum for the purposes of social policy, because he is largely formed in and by the social process, and the nature of the individual must be affected by any social action. Consequently, social policy must be judged by the kind of individuals that are produced by or under it, and not merely by the type of relations which subsist among individuals taken as they stand. (Knight 1947, p. 84)

All social contract theories place agents in a hypothetical initial contractual situation. They do this because they see the right way to answer the question "What sort

[2] Harsanyi's argument is not in fact an argument for utilitarianism at all, though there is still a great deal of confusion regarding this fact. For an especially clear discussion of this point, see (Roemer 1994, ch. 4). Mathias Risse has recently attempted to argue that Harsanyi's social welfare function is indeed a utilitarian one. See Risse (2002). But the argument fails completely, and Risse seems not even to understand the point of Roemer's discussion. See "Section VI: Interpersonal Comparisons of Utility" in (Risse 2002). There are genuine utilitarian theorems—proofs that, given some small set of assumptions regarding the type of social welfare function that may be chosen, only a utilitarian social welfare function will do (d'Aspremont and Gevers 1975; Maskin 1978).

[3] For an early and illuminating explanation of why this is so, see (Wolff 1977).

of society is just?" as by answering the question "What would individual agents in the right sort of setting consent to?" where "consent" is an act which is free, voluntary, and (in some sense) informed. The reason why they believe this second to be the right question is that they think justice in society is a matter of just relations between individual agents. The fundamental error of individual liberalism, which all social contract theories run the risk of making, is to take the individual to be an extra-social entity, and to use the features of this extra-social entity—which are simply posited—to determine the content of the contract such entities would reach, according to which society should be structured. But there is no such thing as an extra-social individual. Therefore, although social contract theories are free to posit whatever individuals they like, they must be recognized as essentially incomplete. A social contract theory requires an antecedent, morally plausible view of what type of agents we have reason to be. With such a view assumed, the theory can then proceeds to ask what sort of society agents of the type we have reason to be would consent to, if they had the opportunity to come to a consensus about their society in an appropriate initial situation. We are then entitled to ask at least two questions of any social contract theory. First, does it indeed have a morally plausible view of the type of agents we have reason to be? If not, the fact that the agents in the theory would consent, given the initial situation, to a certain type of society is not an argument for adopting that type of society. Second, is the society consented to one which is likely to produce individual agents of the type we have reason to be? If not, then the fact that agents of the type we have reason to be would consent to it in the initial situation is of theoretical interest, but no more than this. This is true irrespective of how attractive the initial situation is. It is irrelevant to *our practical question* of what sort of society we should have. The sort of society we should have is one in which we are likely to develop into the type of agents we have reason to be. (There is a separate question of whether the legitimacy of the sort of society we should have can be justified—i.e. whether it can be shown that individuals have a moral duty to support the establishment of such a society, that this establishment does not violate their moral rights. One might think that this case cannot be made without a social contract argument. But I show in Part IV that it can.)

Hobbesian contractarian theories fail these tests, for a number of reasons. There is no basis whatsoever for thinking that the system of social organization that would be chosen by the types of agents imagined by these theories would be one which was capable of cultivating rationality, as conceived of by these theories, in all members of society. The kinds of agent which these theories portray us as—rational in the neoclassical economic sense of rationality—are not the kinds of agents we have reason to want to be.[4] It is doubtful whether human psychology is such that it is even possible for human beings to mimic the behavior of the neoclassical economic agent.[5] Human beings did not evolve to be self-regarding creatures. We evolved to

[4] There are a number of excellent discussions of the problems inherent in the neoclassical economic conception of rationality, and I feel no need to add to them here. See, for example, (Hollis and Nell 1975; Hodgson 2000).

[5] See, for example, (Kahneman et al. 1982).

be reciprocal altruists—a point to which we will return in Chap. 14. Furthermore, Hobbesian contractarianism interprets the notion of a just social arrangement as a solution to a Nash-bargaining problem. This requires ignoring the problems of innate inequalities in bargaining skill, preferences formed under conditions of inequality of opportunity, and inequalities in resource endowments. None of these factors can be ignored by even a remotely plausible theory of social justice (Roemer 1994, ch. 3–4).[6]

Evaluating Nozick's Lockean Contractarianism is a more complicated task.[7] The first difficulty is in interpreting the agents in the state of nature imagined by Nozick. He would, presumably, want them to be taken as rational in the neoclassical economic sense. His argument that these agents would not consent to the formation of anything more than a minimal State, then, must assume that State intervention in the marketplace can never lead to Pareto-improvements, in which everyone is made better off without anyone being made worse off. As we will see in Chap. 10, this assumption is false, given common real-world market imperfections. In order to maintain the argument for the minimal State in spite of this, Nozick's agents would have to be interpreted in a very different way. They are distinctively libertarian agents, each of whose sole interest is in retaining his control over himself and his property, even if this means that everyone, including himself, is thereby worse off. Establishing the minimal State does precisely this. It protects each person's possession of himself and his property from violence, theft and fraud. It also protects the legitimacy of that possession, by affording the same protection to everyone, regardless of their ability to pay for their share of that protection in the form of taxes. It is by extending that protection to everyone (and thus going beyond what Nozick calls the "ultra-minimal" State) that the minimal State creates a presumption that when one acquires additional property through a voluntary transaction, the property one acquires was not previously acquired through violence, theft, or fraud.

---

[6] Ken Binmore has developed a sort of Rawlsian contractarianism, according to which a set of cultural rules similar to Rawls' principles of justice are arrived at through Nash-bargaining in a series of repeated games by essentially self-regarding (but not utility-maximizing) individuals whose behavior is, for evolutionary reasons, irrational (i.e. non-self-regarding) in certain types of situations (Binmore 2005). In addition to the problems that beset any Nash-bargaining approach to social justice, this theory has its own particular difficulty. The argument for the emergence of social cooperation among self-regarding individuals in repeated games depends on a game-theoretic result called the Folk Theorem, which requires very strong and empirically implausible assumptions (Gintis 2005).

[7] One might doubt whether Nozick is a contractarian at all, since his goal is to show that, even given a robust and inviolable right of self-ownership, a minimal State can be established without violating any individual's rights, even if some individualist anarchists never give their consent to it. But the argument can only get off the ground if enough individuals come together to form a protective association which dominates the protection of individual rights in their geographical area. If no one is willing to give up his right to protect himself and his property himself, through the private use of force—if everyone is an individualist anarchist—then no State can emerge. Nozick must assume that the individual incentives to join a protective association are sufficient to set us down the path to the minimal State. So he holds that the State does, and must, begin with a social contract of a sort.

Protecting the legitimacy of possession through redistributive taxation, moreover, does not violate the property rights of the taxed even if they do not wish to pay for others' protection. The State prohibits those who cannot afford to pay for protection from protecting themselves, through the private use of force, from those who can pay for protection. There is an inherent risk that private force will be disproportionate, and will thus violate the rights of those against whom it is used. Because of this inherent risk, the State does not violate the rights of those who would rely on private force when it prohibits them from doing so. Those who can afford to pay the State for protection are not disadvantaged by the prohibition. They become supporters of the State, as this is now the only way to obtain protection. But those who are disadvantaged by the prohibition—those who cannot afford to pay the State for protection—are owed compensation. Redistributive taxation, in order to pay for State protection for those disadvantaged by the prohibition against using private force, is fair compensation. Those who pay the State for protection are morally obligated to provide it. Thus does the minimal State emerge.

Nozick's Lockean contractarianism may pass the second test for avoiding the fundamental error of liberal individualism—perhaps the society he envisions would produce the libertarian agents he takes as given. I doubt it passes the first—but the unappealing nature of Nozick's agents is not a point I wish to argue. Nozick would likely deny that the fundamental error is an error at all; his endorsement of the state of nature tradition in political philosophy implies that he believes that there is such a thing as the natural state of man, which can and should be taken as given, and that this is the state of being a libertarian agent. In Chap. 10, so as not to beg this question against him (outlandish as I think this assumption is), I will undermine Nozick's position by other means. First, I will argue that the minimal State cannot emerge without violating the rights (as Nozick understands them) of libertarian agents. Then, I will counter his argument that the only legitimate State is the one that can be established without violating any agent's moral right of self-ownership, as this right is interpreted by Nozick. There are good reasons to doubt the existence of such a right.

Contractualist theories, in their choice-theoretic variant, suffer many of the same problems as Hobbesian contractarianism. These differ from contractarian theories in only one respect: the agents forming the social contract are assumed to be ignorant of their own particular interests, social standing, and conception of the good life, so that the decisions they make regarding what sort of system of organization to endorse are (supposedly) unbiased and impartial. There is no firmer basis for thinking that the systems of organization derived from such theories will cultivate agents who are rational in the way that the theory assumes the participants in the social contract are, than there was in the case of contractarian theories. And these agents themselves have become no more appealing.

Kantian Contractualist theories at least have the advantage of presenting us with a far more appealing type of agent. As Rawls puts it, the agent who occupies the original position, which is characterized by the same sort of ignorance of personal position found in choice-theoretic contractualism, is one who has

> a capacity for a sense of justice and a capacity for a conception of the good. A sense of justice is the capacity to understand, to apply, and to act from the public conception of justice…the capacity for a conception of the good is the capacity to form, to revise, and rationally to pursue a conception of one's own rational advantage or good…the basic idea is that in virtue of their two moral powers . . . persons are free. Their having these powers to the requisite minimum degree to be fully cooperating members of society makes persons equal. (Rawls 1987, pp. 18–19)

This is not the self-regarding utility-maximizer of neoclassical economic theory. This agent possesses a richer sort of rationality. His preferences are based on reasons. He is instrumentally rational, in the sense that he is capable of determining the best way of satisfying his own preferences, and is interested in their satisfaction. But he also has values and commitments which constrain his pursuit of preference-satisfaction, and he is capable of regarding himself and his fellow citizens as subject to duties which apply independent of anyone's individual conception of a good life. He also possesses communicative rationality: he participates in reasoned debate, and is capable of working out a set of public duties with his fellow citizens, and of being convinced that he ought to endorse one system of social organization rather than another. There is no way to get a Rawlsian social contract out of neoclassical rational agents. There is no Rawlsian social choice theorem—no demonstration that, given some small set of plausible restrictions on the type of social welfare function we may choose, we must choose one that captures Rawls' principles of justice. And so the later, more thorough-going Kantian Rawls attempts a less formal, philosophical argument that his principles are the ones that free and autonomous agents would choose from behind a veil of ignorance.

The question of whether Rawls' theory of justice is adequate—whether it avoids the fundamental error of liberal individualism—then turns on whether a society organized according to Rawls' principles is indeed one which is likely to produce agents who are autonomous and free. In Chap. 9, I will argue that it is not. Little of this argument will be original; the essential points have all already been made by others. I will simply review the most important of the anti-Rawlsian arguments—G.A. Cohen's in particular—and signal my agreement. This is not to dispute Rawls' claim that the actual members of a society organized according to his principles would not endorse those principles, and thus give stability to the system (although as Cohen has argued, they might very well not, at least not under Rawls' own interpretation of them). Rather, what matters is that their actual endorsement would not be the endorsement of genuinely free and autonomous agents.

At that point, it would remain open to one committed to Kantian Contractualism to argue that Rawls' principles are not the ones that autonomous and free agents would choose; rather, they would choose to structure their society in a way which was, at a minimum, suited to producing agents like them. This move, however, has the potential to undermine the social contract approach to justice. One way to interpret it is as establishing a goal: whatever else it does, an adequate theory of justice must tell us how society must be organized in order to produce free and autonomous members. If this is the appropriate goal of a theory of social justice, then what we ought to do is figure out how to achieve it; what is needed is an argument that a

particular set of principles, or institutions, or policies, is what is required in order to meet this goal. The legitimacy of the resultant theory will derive from the fact that the theory solves the problem; it identifies the required principles, institutions or policies, and provides an argument connecting them with the achievement of the goal. The whole idea of a social contract has now been written out of the story. It is no longer needed to establish the legitimacy of the resulting theory of social justice, and it plays no methodological role in determining the content of the theory. We may of course believe that, after the fact, the free and autonomous citizens produced by such a society would endorse the system of social organization that succeeded in producing them. But that is no comfort to the Kantian contractualist. For his approach to survive, it must be possible to design the initial contractual situation in such a way as to yield this result, *without* imposing it from the beginning as the goal of the contract making. And it must be possible to justify the design of this initial contractual situation, and argue that the legitimacy of the resulting system of social organization derives from it, *without* appealing to this goal. There is no Kantian contractualist theory in the offing that succeeds in doing this.

Ideal speech situation theory, on the other hand, looks far more promising. This variety of contractualism identifies the initial contractual situation as one of equal power among the participants in the contractual negotiations. There is neither coercion nor manipulation; every participant has an equal opportunity to express his views, and to question the views of others; and every participant holds a veto. The assumptions made about the agents are much weaker. They are fairly ordinary individuals possessed of communicative rationality, the capacity to exchange reasons in debate and to be convinced by the position of another when the reasons that favor it are seen to be stronger. This makes it much easier for this type of theory to avoid the fundamental error of liberal individualism. There is neither anything especially admirable nor anything unappealing about its agents. There is no veil of ignorance; individuals know their own interests and values. One of the major criticisms—and a perfectly valid one—made by ideal speech situation contractualists against Rawls' theory is that the veil of ignorance makes the very idea of negotiation among distinct individuals, which is the core of social contract theory, inert. Debate and negotiation are based on differences of interests, values, and points of view. The veil of ignorance obscures all such differences.

In Anglo-American political philosophy, the dominant form of ideal speech situation contractualism is the Scanlonian variety—developed by T.M. Scanlon and Brian Barry (Scanlon 1998; Barry 1995). This version of the theory will not detain us. Its defining feature is the claim that, if the participants in the social contract negotiation are only permitted to use their vetoes to reject proposals which are "reasonably rejectable," they will eventually reach a consensus around a set of organizing social principles which cannot be reasonably rejected. These are the principles of justice, and their normative status derives from the fact that communicatively rational agents in a position of equal power could not reasonably reject them. As Daniel Bell has pointed out, it is terribly difficult to know what to make of any of this, since the Scanlonians have precious little to say about what it means for a pro-

posal to be reasonably rejectable or not; what makes one so; or how to determine whether one is so. They offer instead a list of the sorts of considerations which they do not consider appropriate grounds for reasonable rejection—such as conflict with religious belief—while offering little in the way of defense of the items on the list (Bell 1998, pp. 563–564). Moreover, they offer no theoretical basis for the expectation that this process of debate will ever end in universal consensus around a set of organizing social principles—just as Rawls fails to provide any such basis for his expectation that the process of reflective equilibrium, whereby conclusions reached behind the veil of ignorance are tested against the values held by the participants when not behind the veil and then revised, will eventually come to an end (Bonevac 2004).

Gerald Gaus has recently developed his own detailed version of ideal speech-situation contractualism (Gaus 2011). Gaus' contracting agents are hypothetical idealizations, but they are far more realistic than the agents behind Rawls' veil of ignorance, and there is no reason to think that Gaus' theory fails to avoid the fundamental error of liberal individualism. Nonetheless, we need not dwell on the specifics of Gaus' view. For he does not manage to avoid the deep problems that plague Rawls, or indeed any version of contractualism which grounds the legitimacy of social organization on a hypothetical agreement among idealized agents. The following passage is especially telling:

> As an actual person participating in social morality, when I make a demand on you to conform to a publicly justified rule [i.e. a rule which Gaus' idealized agents would consent to], I am claiming that I have standing to direct your actions because your own reason has accorded me that standing, though you are now failing to see, or refuse to concede, that you must conform to the rule. (Gaus 2011, p. 29)

The problem here is obvious. Gaus identifies consent as the only legitimate basis for authority. To put it in his own terms, he holds that regardless of what reasons there may *be* for someone to act, he can only be held accountable for the reasons he *has*—the reasons he recognizes as applying to him (Gaus 2011, pp. 232–235). There is nothing wrong with others working to convince him that a given reason for action does apply to him. But they can only hold him accountable for not acting on it if they succeed. What Gaus does not, and cannot, defensibly maintain is that consent is the only legitimate basis for authority, but that hypothetical consent is just as good as actual consent. There is no acceptable move from a consent-theory of authority to the conclusion that we are justified in holding others responsible for what idealized versions of themselves *would* consent to in idealized conditions, even though *they have not*. And yet this is precisely what Gaus advocates. He attempts to make an argument by analogy with promising, and points out that there is nothing problematic about holding someone to a promise he has made, even if he decided later on he does not wish to live up to it (Gaus 2011, p. 30). But in Gaus' vision of society, the actual agents have not committed themselves to anything. There is no actual social contract. There is only the hypothetical contract which Gaus argues would be made by suitably idealized agents. But what is left of the claim that

authority can only derive from consent, when we immediately turn around and deny that any actual consenting must be done by any actual people?[8]

Gaus might want to argue (as I interpret Rawls as arguing) that hypothetical consent is not just as good as actual consent, it is *better*. That is, he might try to argue that the obligations that should be imposed on and exacted from a person are, constitutively, those that a hypothetical free, rational, moral version of himself would consent to, and that the legitimacy of such obligations is in no way derived from or parasitic on the apparent legitimacy conferred by actual consent. In this case, he would be defining—or rather, re-defining—"the reasons one has" as "the reasons an idealized version of oneself would have in idealized circumstances." So long as the idealization is not too strict, there may still be room to distinguish between the reasons one has and the reasons there are. Perhaps this is how we should read Gaus—I will admit to some uncertainty on this point. But if so, then his project (unlike Rawls', I believe) is self-defeating. For given Gaus' understanding of authoritarianism, and the fact that producing a non-authoritarian theory is a central goal of his project, Gaus cannot completely untether himself from an actual consent-based theory of authority.

What Gaus fails to see is that just as one actual agent may disagree with another about what he is obliged to do, we may very well disagree with Gaus about the content of the supposedly authority-conferring hypothetical social contract. Here Gaus makes a mistake also made by Rawls: rather than merely arguing that consensus following debate in an ideal contractual situation is the only acceptable way of arriving at principles of social justice, Gaus goes on to argue that we can determine what principles that consensus would form around, without having to go through the trouble of even approximating this deliberative exercise. The consequences of this mistake for Gaus' theory are even more damaging than they are for Rawls'. Suppose that Gaus tries to enforce some supposed obligation of social morality on me, and cites as the source of his standing to do so the fact that I am rationally committed to recognizing this obligation, even if I fail/refuse to do so now. The point he seems to be oblivious to is that I may very well disagree, not only with his claim that I have this obligation, but also with his claim that this is one of the obligations my free, rational, moral self would voluntarily take on. For I may have my own ideas about what this version of myself would and would not agree to, and what a community of these selves would and would not form a consensus around. Or I may hold that it is preposterous for anyone to think that he or she has come up with a definitive argument as to what the content of this consensus would be, and abstain from assenting to any view on the matter. And since this burdening of myself is a hypothetical burdening of a hypothetical version of myself, there is no fact, no event like the making of a promise, that Gaus can point out to show me that I am committed in the way he claims. His only recourse at this point is to claim that I am in the dark about what agreements a community of free, rational and moral selves would reach, and that I should re-read his work as many times as I need to in order to become convinced.

[8] David Enoch pursues this familiar line of thought with admirable thoroughness (Enoch 2005).

The upshot of this is that Gaus may be able to claim that I have such-and-such obligation not because he says I do, but because I myself am rationally committed to accepting that obligation. But if I dispute that I am rationally so committed, and doubt or contest his arguments that I am, his only possible response is that his arguments do show that I am, whatever I may think about the matter. If it really were authoritarian to hold someone accountable to reasons he does not recognize, as Gaus claims, then he would escape authoritarianism at the level of specific obligations only by embracing authoritarianism (as he understands it) at the level of the hypothetical commitments that supposedly ground those obligations. In order for Gaus' exercise of authority over me to be non-authoritarian by his own lights, I would have at least had to buy into his account of what the content of the hypothetical social contract would be, even if I did not presently agree that I have the specific obligation he is pressing me on. That is to say, I would have to *actually consent* to something at some point in this story, however far up the chain of abstraction that might be, in order to provide an anchoring point from which my commitment to fulfilling some specific obligation could be demonstrated. Gaus cannot treat himself to hypotheticals all the way up.

As we will see in Chap. 15, I do not subscribe to a consent-based theory of practical authority. Gaus' contention that it is authoritarian to hold another person accountable for fulfilling an obligation he himself does not recognize is, as far as I am concerned, an anarchical fallacy.[9] And so I need not be troubled by Gaus' failure to ground the legitimacy of his system of social organization in his preferred way. There remains the option of evaluating the system of social organization Gaus generates, freed from the background against which he tries to derive its legitimacy. Gaus' classical liberalism, however, while having much in common with Rawls' liberalism, casts the State in an even smaller role than Rawls does. And as I will argue that Rawls' system of social organization falls short as a defensible articulation of the goal of social justice, I may be taken as arguing *a fortiori* that Gaus' does as well.

The neo-Kantian contractualism of Habermas, on the other hand, is a form of ideal speech situation theory worth taking very seriously. It is the only theory of this type which is properly grounded. Habermas develops a theory of communicative rationality, which serves as the right sort of foundation for the claim that the participants in the ideal speech situation must eventually reach a consensus. He begins from a particular theory of linguistic meaning and the very conditions of meaningful communication (Habermas 2000). He thus takes us out of the Kantian world of individual rationality and relocates us in the social, intersubjective world of communicative rationality. He then argues that communicatively rational participants to moral discourse who occupy an ideal speech situation—one in which they occupy equal power, and are free from time constraints and other mundane limitations—would reach universal, deliberative consensus on moral principles in virtue of their adherence to the principles that underlie the very possibility of communication (Habermas 2001). Habermas, like Scanlon, accepts a norm as justified if, and only

[9] See the first of William Edmundson's three superb essays in (Edmundson 1998).

if, it is not reasonably rejectable by anyone who would be affected by observance of the norm. But his theory has the resources to do what Scanlon's cannot: it can give a precise definition of what it means for a proposal to be reasonably rejectable. A proposal is reasonably rejectable so long as an argument for its rejection can be stated and defended from all objections made in accordance with the principles of communication in an ideal speech situation, without transgressing those same principles.

I share Habermas' hope for a stable convergence of normative judgment—a commonality that is largely due to the shared influence of C. S. Peirce. I do not accept his argument for believing that such a convergence must be possible at least in principle, or the highly controversial theory of linguistic meaning and communication on which that argument depends.[10] My own interpretation of this Peircean hope, and my own reasons for maintaining it, will come in Chap. 13. But even if Habermas' views on all these points are correct, I do not think that he succeeds in translating his neo-Kantian contractualism into a foundation for an adequate theory of social justice. Habermas (quite rightly) does not assume that it is possible to determine, *ex-ante*, what the conclusions of participants in anything approximating an ideal speech situation would be. So he does not presume to claim—as Rawls and Gaus do—that he can lay down principles of social justice which would receive the endorsement of individuals so situated. And he is well aware that the world as it is is a far cry from being even a reasonable approximation to the ideal speech situation required for the discovery of normative truths. The political program that Habermas' neo-Kantian contractualism yields, therefore, is that of moving society as far in the direction of becoming an ideal speech situation as is possible—knowing all the while that a perfect instantiation is impossible, since the ideal speech situation is an idealization which transgresses the boundaries of actual human life. Habermas' theory thus takes seriously one of the major problems that confounds Rawlsian and Gausian contractualism. It is not, strictly speaking, a social contract theory at all. Rather, it is a goal-directed theory, where the goal is to come as close as possible to realizing an ideal speech situation, in which the sort of debate which is capable of leading to normative consensus will finally become possible. It is thus a precursor theory to a future contractualism—a theory which sets as its goal the creation of the right kind of contractual situation.

Habermas' particular politico-legal interpretation of what it would mean to move society in the direction of realizing an ideal speech-situation is moving society in the direction of deliberative democracy—a system of social organization in which political decisions and choices of social and economic policy are made on the basis of extensive, sincere and informed public debate, aimed at establishing broad-based

[10] For a sympathetic and highly informed critique of Habermas' pragmatics, discourse ethics, and the link between the two, see (Heath 2001). My own primary objection to Habermas' meta-ethical and normative ethical theory is that he does not see moral claims as based on evidence and answerable to the world (Habermas 1998). I obviously do see them in this way, though not as based on the same sort of evidence or answerable to the world in precisely the same way as scientific claims. I return to this point in Chap. 13.

consensus, among the greatest possible number of participants, all of whom occupy positions of equal power. Once a true deliberative democracy was established—or at least the best possible approximation to one—then the genuine principles of social justice could be discovered. The best argument against the claim that deliberative democracy is a necessary precursor to social justice is still Benjamin Constant's (Constant 1819/1988). Deliberative democracy requires that every member of the political community exercise what Constant called "the Liberty of the Ancients": direct and continuous participation in considerations of political matters, public debate, and social decision-making. As Constant observes, this way of life—and it is an entire way of life—is only feasible if society is small to begin with; if there are severe restrictions on citizenship, so that the number of political participants is kept to a manageable size; if a great number of diverse systems of value are not found among the citizenry; and if there is a non-citizen servant class whose labor supports the politically engaged lifestyle of the citizenry. It is simply not possible, in other words, in the large, diverse, modern, industrial nation-state. We have good reason to believe, then, that perfect deliberative democracy is the wrong politico-legal interpretation of the ideal speech situation, because we have good reason to believe that it is *not* the system of social organization which would be chosen by the members of a large, diverse, industrial society in an ideal speech-situation (without, of course, claiming to know in advance what choice they would make).

2 Goal-Directed Theories

The other two prominent goal-directed theories which compete with egalitarianism are utilitarianism (or more generally, any aggregate value maximizing view) and Derek Parfit's prioritarianism. We have already seen the first of the criteria for any adequate theory of distributive justice:

(1) Any theory of distributive justice must avoid the fundamental error of liberal individualism.

I now introduce two others:

(2) Any theory of distributive justice must advocate a non-exploitative scheme of social organization.
(3) Any theory of distributive justice must respect autonomously chosen effort and risk with respect to the distribution of well-being.

Criterion (2) is to be interpreted using John Roemer's account of what it means for one group of agents to *exploit* another (Roemer 1982, pp. 194–195):

Suppose society is divided into the two groups C and W. Group C exploits group W if

(1) There is an alternative social arrangement, which we may conceive of as hypothetically feasible, in which group W would be better off than in its present situation.

(2) Under this alternative group *C* would be worse off than at present.
(3) Group *C* is in a relationship of dominance to group *W*. This dominance allows it to prevent group *W* from realizing this alternative.

The relevant sort of alternative social arrangement is one in which the members of group *W* withdraw from society in accordance with a withdrawal rule. By accepting the first two of these criteria, we are able to turn the social contract approach to distributive justice on its head. Rather than simply positing a certain type of individual (whatever type the theory identifies as appropriate for making a social contract), and trying to deduce what contract would be made by individuals of that type in a given (also allegedly appropriate) contractual situation, we take seriously the fact that the individual is profoundly influenced by his social, cultural, political and economic environment, and explicitly begin with the goal of determining what type of individuals we have reason to want to become. The first part of our task is then to find a way of organizing society that is likely to produce individuals like that. The second part is to further refine that system of organization so that no group of individuals would be better-off withdrawing from society. Thus, rather than assuming that it is possible to design an initial contractual situation in which the sort of individuals we have reason to become would agree to *enter into* a society capable of producing agents like them, we instead try to find a way of organizing society which is capable of producing individuals like them, and which none of those individuals would want to *withdraw from*. This reversal leaves open the question of legitimacy, since a society no one would opt out of is not necessarily a legitimate one, and in fact adds to the burden of justifying legitimacy by demanding more of our conception of the sort of society we should have, and thus serving as a source of additional or strengthened duties on individuals. But again, the justification of the legitimacy of a just society will come in Part IV. And, given that its legitimacy can be established, there is no reason to believe that a society from which no one would withdraw would be less stable than a society to which hypothetical versions of ourselves would arguably consent. The third criterion, finally, is needed to ensure that, in a non-exploitative society, the most talented and productive members do not withdraw.

My argument against utilitarianism is that it does not satisfy the second of these criteria. That it does not will become clear from my discussions of these requirements in the chapters that follow. I am fully aware that this is not a definitive argument against utilitarianism. I do not think that there is any such argument. The utilitarian is free to dispute the notion that these last two criteria are requirements of any adequate distributive theory. What allows the utilitarian to do this is the fact that his theory rests on a commitment which is arguably even deeper than any of the requirements discussed so far. This is the fundamental challenge that motivates all forms of utilitarianism:

(U) We cannot justify doing less good than we could do.

The key to understanding this challenge, as stated by the utilitarian, is that "no less good than we could do" means the maximum aggregate of some interpersonally

comparable value. If all we mean by the best social arrangement is the one we should most prefer, then the advocate of an opposing theory—the egalitarian, say—is free to contend that his theory, if put fully into practice, would be best—would realize the greatest possible good—even though the utilitarian would disagree. He is free to employ a different way of reckoning which of any two possible social states is better—which realizes greater overall goodness or value—than would a utilitarian. I do in fact think that the society of Equal Liberty is the best possible society, the one that should be placed first in any ranking of possible social systems by overall realization of goodness or value, even though it may not be the society that maximizes aggregate interpersonally comparable welfare (where I would understand "welfare" as level of achieved functioning, following Sen). Of course, the utilitarian could not agree, and I would not expect him to. But what I cannot, and do not, claim is that we ought to have creating a society of Equal Liberty as our overarching social goal *in virtue of* the fact that it maximizes the aggregate of any compositional value (even if it does do so). This is what makes my view non-utilitarian. When the egalitarian asserts that the society of Equal Liberty is best, is the one that realizes the greatest overall goodness or value, he could just as well say that it is, for some further reason, the society we ought to aim for, all things considered. What is that further reason, which is the root of any more precise egalitarian criterion for ranking social states? Ultimately, the conflict between utilitarianism and egalitarianism derives from the fact that they are motivated by distinct, but equally profound, challenges. The thought that motivates egalitarianism is the following:

(E) We cannot justify allowing some, through no choice or fault of their own, to fare worse than others.

My allegiance to egalitarianism, at bottom, derives from the fact that I am more greatly moved by the challenge that motivates it than I am by the challenge that motivates utilitarianism. There is, as I have said, no definitive argument to be had here.

There is no similarly deep challenge motivating prioritarianism—it is not a primordial theory of distributive justice, as utilitarianism and egalitarianism are. Rather, it is a view which was designed to preserve the attractive features of these views while avoiding their perceived excesses. My argument against prioritarianism will largely take the form of a demonstration that my brand of egalitarianism does not suffer from any of the faults which prioritarians have charged egalitarianism with. In each case in which it is alleged that prioritarianism yields a more appealing conclusion than egalitarianism, I will argue that, at least as far as the theory of Equal Liberty is concerned, the opposite is the case. Since prioritarianism lacks any deep independent motivation, exposing its failure to identify any serious flaws in my form of egalitarianism, or to deal better with any proposed distributive problem, undermines the case for it.

Once the path of egalitarianism has been cleared, we must turn our attention to the question of which member of this body of theories to endorse. We must confront the question of what quality should be equalized across the members of society. The

theory of Equal Liberty provides a precise answer to this question. Over the course of the next three chapters, I develop the theory in detail, and provide both negative arguments—critiques of what I take to be its most powerful egalitarian competitors—and positive ones on its behalf.

References

Barry, B. 1995. *Justice as impartiality*. Oxford: Clarendon Press.
Bell, D.A. 1998. The limits of liberal justice. *Political Theory* 26(4): 557–582.
Binmore, K. 2005. *Natural justice*. Oxford: Oxford University Press.
Bonevac, D. 2004. Reflection without equilibrium. *The Journal of Philosophy* 101(7): 363–388.
Cohen, G.A. 1995. *Self-ownership, freedom, and equality*. Cambridge: Cambridge University Press.
Constant, B. 1819/1988. The liberty of the ancients compared with that of the moderns. In *The political writing of Benjamin constant*, ed. B. Fontana, 309–328. Cambridge: Cambridge University Press.
d'Aspremont, C., and L. Gevers. 1975. Equity and the Informational Basis of Collective Choice Faculté des Sciences Économiques et Sociales, Working Paper 7–5.
Edmundson, W.A. 1998. *Three anarchical fallacies*. Cambridge: Cambridge University Press.
Enoch, D. 2005. Why idealize? *Ethics* 115: 759–787.
Gaus, G. 2011. *The order of public reason*. Cambridge: Cambridge University Press.
Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.
Gintis, H. 2005. Behavioral ethics meets natural justice. *Politics, Philosophy, and Economics* 5(1): 5–32.
Habermas, J. 1998. Rightness versus truth: On the sense of normative validity in moral judgments and norms. In *Truth and justification*, 237–276. Cambridge, MA: MIT Press.
Habermas, J. 2000. Universal pragmatics. In *On the pragmatics of communication*. Cambridge, MA: MIT Press.
Habermas, J. 2001. *Moral consciousness and communicative action*. Cambridge, MA: MIT Press.
Heath, J. 2001. *Rational choice and communicative action*. Cambridge, MA: MIT Press.
Hodgson, B. 2000. *Economics as a moral science*. Berlin: Springer.
Hollis, M., and E. Nell. 1975. *Rational economic man*. Cambridge: Cambridge University Press.
Kahneman, D., P. Slovic, and A. Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.
Knight, F.H. 1947. Ethics and economic reform. In *Freedom and reform: Essays in economics and social philosophy*, 55–153. New York: Harper.
Maskin, E. 1978. A theorem on utilitarianism. *Review of Economic Studies* 45(1): 93–96.
Nozick, R. 1974. *Anarchy, state and utopia*. Cambridge, MA: Harvard University Press.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Rawls, J. 1987. *Political liberalism*. Cambridge, MA: Harvard University Press.
Rawls, J. 2001. *Justice as fairness: A restatement*. Cambridge, MA: Harvard University Press.
Risse, M. 2002. Harsanyi's 'utilitarian theorem' and utilitarianism. *Noûs* 36(4): 550–577.
Roemer, J. 1982. *A general theory of exploitation and class*. Cambridge, MA: Harvard University Press.
Roemer, J. 1994. *Theories of distributive justice*. Cambridge, MA: Harvard University Press.
Scanlon, T.M. 1998. *What we owe to each other*. Cambridge, MA: Harvard University Press.
Wolff, R.P. 1977. *Understanding rawls: A reconstruction and critique of a theory of justice*. Princeton: Princeton University Press.

Chapter 9
Liberty, Equality and Justice

1 Introduction

The work of the last several chapters has given us a rigorous account of autonomy, a precise and measurable conception of freedom, and a compound conception of liberty that integrates the two. The aim of this chapter and the next three is to develop the essential elements of a theory of distributive justice that takes liberty as the focus of the State's distributive concern. As will become clear below, the theory developed here is faithful to the basic ethos of egalitarianism, and so I will refer to it as the "Equal Liberty" theory of distributive justice. The question of the ground of the State's authority to pursue such a scheme of distributive justice will be the subject of Part IV of the book.

This chapter will proceed for the most part as a discussion with, and an interrogation of, a series of interlocutors representing rival distributive theories. The views considered include Derek Parfit's recently developed prioritarianism (an attempt to solve the distributive problems of utilitarianism in a way that improves on strict egalitarianism); Martha Nussbaum's satisficing egalitarianism; and Ronald Dworkin's resource-based egalitarianism. I also give a fairly cursory treatment of Rawls' vast theory of Justice as Fairness. I do not believe that my approach to distributive justice opens the way to any objections to Rawls' view that have not been made before. My discussion of Rawls, then, will be brief and largely limited to registering agreement with objections that have already been made and that are consistent with and supported by my own approach. One conspicuous absence from this part of the book is a discussion of Rawls' theory of autonomy. I address this aspect of Rawls' thought in Chap. 16 in the context of a critical examination of Jonathan Quong's argument for non-perfectionist liberalism, in which adherence to a Rawlsian conception of autonomy plays a central role. I then begin constructing my own positive account with some ground-clearing. I begin with a simple statement of the basic egalitarian ethos, and defend its appropriateness as an expression of pure social justice. I frame the development of the Equal Liberty account as an attempt to flesh out the basic idea behind egalitarianism in a way that avoids the

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_9

pitfalls of the other views discussed. As I develop my account in this chapter and those that follow, I will introduce and respond to numerous objections that have been raised by Richard Arneson and others against opportunity-based egalitarianism—a family of theories of distributive justice to which Equal Liberty belongs.

2 Theories, Ideal and Non-ideal

It is currently popular to ask of a theory of distributive justice whether it is an "ideal theory" or a "non-ideal theory." This question, however, lacks the clarity needed for its answer to be interesting and significant. In this section I briefly articulate what I take to be the distinction between an ideal and a non-ideal theory of justice (or rather the distinctions between various ways in which a theory may qualify as ideal or non-ideal), and locate my own theory on this spectrum.

The best way to approach the issue of ideal vs. non-ideal theories is to begin by specifying the elements of a comprehensive theory of distributive justice—an exercise which is also useful in developing any such theory, regardless of which side of this divide it falls on. At the most basic level, there are three such elements:

1. *The aspirational account*: This is the picture of what a just society looks like, according to the theory. It describes the just distribution and the effects of that distribution on the lives of the members of the society. It includes both (a) maintenance principles, which guide the return to the aspirational scenario after a departure caused by an exogenous shock; and (b) preservation principles, which ensure that the aspirational scenario will continue to be realized in future generations.
2. *The descriptive account*: This is a model of the current, less-than-just system of social organization, including the principles that maintain and preserve that system.
3. *The transitional account*: This is the statement of the principles of distribution and social organization that society ought to adopt in order to begin transforming itself into a just society.

We can now identify four types of theories along the ideal/non-ideal spectrum according to the way they characterize each of these elements. The first type, which I will call "strong ideal theory," constructs an aspirational account which is insensitive to normal human psychological and behavioral limitations. The transformation principles offered by such an account are identical to the aspirational principles of maintenance and preservation. An account of this type says, in essence, that a just society looks like *this*, that the way to realize a just society is simply to start behaving in the way that just agents would behave, and that if the members of our society cannot do this because their commitment to justice will never be enough to motivate them to this behavior, then "their very make-up is unjust: they cannot help being unjust" (Cohen 2008, p. 155). In his most recent (and, sadly, final) work, G.A. Cohen has moved toward a theory of this type. Such a theory, he claims, is the only kind

capable of capturing pure justice. Theories which are guided by a recognition of human psychological limitations are not theories of justice *in stricto sensu*, but are instead theories of optimal social regulation. "The impossibility of justice" he asserts, "whether or not it is due to a flaw in human nature, is insufficient for the justice of the possible" (Cohen 2008, p. 155).

The second type of theory, which I will call "weak ideal theory," does take ordinary human psychological limitations into consideration when constructing its aspirational account. Such a theory endorses and is concerned with the justice of the possible. But theories of this second type maintain that the principles that maintain and preserve the aspirational scenario are, if not identical to, then the only appropriate guidelines for, the transitional principles which will transform an unjust society into a just one. On this view, an aspirational account of justice does not only tell us what we are aiming for; it also guides us to that aim. Rawls subscribes to weak ideal theory, asserting that whatever precise transitional principles turn out to be most appropriate for a given society, an outline of an ideal theory of distributive justice is needed in order to formulate them, since it is the ideal principles of justice that must guide the transition.[1]

The third type of theory, which I will call "hybrid theory," differs from weak ideal theory in rejecting the latter's assumption of the relevance of ideal principles to the formulation of transitional principles. A hybrid theory recognizes the possibility that, for a given society, the principles which will best govern the transition from injustice to justice will differ considerably from the principles that will maintain and preserve the aspirational scenario one it has been realized.

The fourth and final type of theory, for which I will reserve the term "non-ideal" theory, claims that we can, and should, devise transitional principles which will move society in the direction of greater justice without first constructing an aspirational account. Such a theory is meant to proceed by identifying sub-optimal outcomes within the current system of social organization, identifying the policy or policies that cause those outcomes, and then altering them so that they no longer cause those undesirable outcomes. The problem with non-ideal theory, as I have defined it, is obvious. In the absence of an aspirational account, there is no basis for robust and specific descriptions of what it is about the undesirable outcomes that make them undesirable; there is no framework in which to pose the question of how the policies that result in undesirable outcomes should be changed; there is no basis for ranking proposed changes. Without an aspirational count to set our aim, there is no basis for formulating a transitional strategy. The reason for this is simple: to debate the merits of different policy choices and courses of action in the absence of a conception of the just society is nothing more than to debate about means without any clear idea of the end being aimed at.

I thus reject non-ideal theory. There is something to be said for strong ideal theory, at the other end of the spectrum. I believe that Cohen is right to maintain that we cannot dismiss *a priori* the possibility that human nature is simply incompatible with the creation and preservation of a truly just society. And I recognize the attempt

[1] See, for example, (J Rawls 1999, pp. 89–90).

to articulate the conditions of pure justice as a legitimate philosophical project. But it is not the project that interests me here. And even if we must draw a distinction between pure justice and "the justice of the possible," I do not think it an abuse of language to refer to a theory concerned with the latter as a theory of justice. At the very least, the claim that we may use the term "justice," and speak of just societies, in both these senses has the greatest of historical pedigrees. Plato's *Republic*, recall, contains descriptions of not one just city-state, but two. The first is the simple *polis*, and Socrates is finished constructing it half-way through Book II (Plato, *Republic*). The great early turning point in that work is the moment when Glaucon responds to Socrates' outline of the just and simple *polis* by pointing out that Socrates "would make these men have their feasts without relishes," and accuses him of having constructed a *polis* fit for pigs.[2] Socrates accepts the objection and goes on to search for justice in the luxurious or "feverish" *polis*. The world we live in is a feverish one and it may well be that there is nothing we can do to permanently change that. If that is so, I side with Socrates, and against Cohen, in thinking that we can, and should, search for the form of justice that is appropriate to our circumstances.

My own approach, then, falls into the category of hybrid theory. I reject weak ideal theory on the grounds that I see no reason to assume that ideal principles are the only appropriate source of guidance for transitional principles. My work in this chapter will in fact focus almost exclusively on the construction of an aspirational account. But I will also discuss the work of John Roemer, who has formulated a social welfare function for an approach to distributive justice he calls "equality of opportunity for welfare." Roemer's account leaves something to be desired if it is taken to be an aspirational account of distributive justice, and some have criticized it assuming that that is what it is supposed to be. Whether Roemer himself conceives of it in that way will not concern me. But I will suggest that Roemer's social choice mechanism may be well suited to play the role of a transitional principle within a theory of distributive justice that includes my aspirational account. As we will see, the principles that maintain and preserve equal liberty are not identical to Roemer's distributive principle. But this is just what we should expect in a hybrid theory. Moreover, the criticisms that have been directed at Roemer's work, criticisms which miss the mark if we assume that Roemer is not engaged in constructing an aspirational account, do not hold against the genuine principles of Equal Liberty.

3 Neutrality, Pluralism and Liberalism

3.1 The Principle of Neutrality and the Perfectionist Critique

The approach to distributive justice which I will develop can be fairly classified as a liberal perfectionist one. It is a liberal theory not only in the literal sense of focusing on liberty, but also in the contemporary sense of being broadly egalitarian—its

[2] See (*Republic*, Book II).

goal is equal liberty, a notion that I will flesh out considerably later on. It is a perfectionist theory insofar as it takes liberty—or at least, the sort of liberty that we should be concerned with morally and politically—to be the liberty to achieve a high level of well-being, and the theory of well-being that has been adopted is a perfectionist one. Well-being is understood as consisting in the pursuit and achievement of valuable goals through the excellent development and exercise of one's abilities and capacities.

For the past 25 years at least, there has been an intense debate over whether a political theory could be both liberal and perfectionist.[3] The debate has centered on the question of whether political liberalism requires a commitment to the Principle of Neutrality.[4] Thanks to the valiant efforts of perfectionists during this debate, the weaknesses of the arguments in favor of neutrality have been exposed. An attempt on my part at an original critique of the pro-neutrality position would add very little to the state of the debate at this point. My first task in this section, then, will merely be to set out as clearly as possible the claims and arguments of the pro-neutrality side, and then briefly discuss the main objections advanced by perfectionists. Having done that, I will proceed to examine some problems which threaten to arise when we abandon neutrality and embrace perfectionism. These problems have not received adequate attention from perfectionists, and original solutions are needed.

The Principle of Neutrality comes in various forms. The two most plausible forms, which we can call the "Intention-neutrality Principle" and the "Argument-neutrality Principle," are as follows:

(A) Intention-neutrality

(i) Citizens are free to accept policies for their own reasons, based on their own comprehensive worldviews.
(ii) Policy-makers are free to address different arguments for policies, tailored to different worldviews, to different groups.
(iii) Policy-makers, however, *must* intend and design their policies exclusively to promote generally shared values.

(B) Argument-neutrality

(i) Citizens are free to accept policies for their own reasons, based on their own comprehensive worldviews.
(ii) Policy-makers are free to intend and design their policies to promote controversial values.
(iii) Policy-makers, however, *must* only adopt policies that can be justified using arguments that contain as premises only generally shared beliefs and value-assertions.

[3] Prominent ant-perfectionists have included Rawls, Dworkin, Brian Barry and Bruce Ackerman. Leading perfectionists include Raz, Thomas Hurka, George Sher and Steven Wall.

[4] Or what Wall calls the Principle of Restraint (Wall 1998).

To see the conflict between neutrality and broadly perfectionist political theories, we can look to Steven Wall's characterization of perfectionism (Wall 1998, p. 7):

1. Political authorities should take an active role in creating and maintaining social conditions that best enable their subjects to lead valuable lives.
2. Some ideals of human flourishing are sound and can be known to be so.
3. The State is presumptively justified in favoring these ideals.
4. A sound account of political morality will be informed by these ideals.
5. There is no general moral principle forbidding the State from favoring these ideals, and enforcing conceptions of political morality informed by them, when these ideals are controversial or subject to reasonable disagreement.

It is the fifth perfectionist claim that conflicts with neutrality. To see this, note that both versions of the neutrality principle forbid policy-makers from adopting certain sorts of policies, even if those policies are among those that would best enable citizens to lead valuable lives. The first neutrality principle excludes any policy designed to favor a sound but controversial ideal, and the second excludes any policy that cannot be justified without recourse to controversial premises, which the justification of perfectionist policies may require.

Obviously claims (A)(iii) and (B)(iii) are the ones that stand in need of argument. Wall has isolated the primary arguments that have been offered by the pro-neutrality side for each of these claims. The primary argument for (A)(iii) is Rawls' argument from the value of political toleration. Closely related to this argument is Ronald Dworkin's argument from the value of equal concern and respect (Dworkin 1985). Rawls' argument can be summarized as follows (Rawls 1987, p. 58):

1. A reasonable person is one who recognizes that others can reasonably disagree with him even when he is correct.
2. We must tolerate reasonable people and reasonable disagreement.
3. To tolerate reasonable people and reasonable disagreement is to adopt policies designed to promote values about which we do not reasonably disagree.
4. So we must only adopt policies designed to promote values about which we do not reasonably disagree.

Wall's objection to this argument, and thus his objection to the first neutrality principle, focuses on the second premise (Wall 1998, p. 79). Toleration, he observes, has a cost. If there is reasonable disagreement about some of those policies that would best enable the members of a society to lead valuable lives, then the cost of Rawls' political toleration is abandoning these policies and losing the benefits to society that they would bring. Rawls' argument seems to require that we assert that the value of toleration always outweighs the costs. But no argument for this very strong claim is forthcoming, and it is highly implausible that toleration, however good it may be, is a special sort of good deserving of the lexical priority this argument must effectively grant it. The same goes for the value of Dworkin's notion of showing equal concern and respect, with which toleration may be replaced in the above argument.

There is an argument for intention-neutrality which does not require the implausible claim that the value of tolerating reasonable disagreement (or of showing equal concern and respect) always outweighs the costs of these actions. One might argue that intention-neutrality is necessarily constitutive of liberalism itself, so that if we abandon these policies for the sake of other social goods, even just occasionally, we will have effectively abandoned liberalism. Insofar as we are committed to liberalism then, we will have to put up with the fact that we are sometimes prevented from promoting the good of society by this very commitment.[5] Such an argument would have to proceed by establishing that without being committed to the Principle of Intention-neutrality, one could not be committed to the Harm Principle:

The Harm Principle: The only adequate justification for state interference in individuals' lives is the prevention of harm in the form of restrictions on other individuals' freedom or autonomy.

Since my theory of the ground of the State's legitimate authority will not be complete until Chap. 15, I will defer addressing this possibility until Chap. 16, where I address the issue of the appropriate limits on that authority. There, I will implicitly undermine this argument by demonstrating the compatibility between political perfectionism and this core liberal principle.

Let us proceed then to the second neutrality principle. The primary argument for (B)(iii) is Rawls' publicity argument (Rawls 1987, pp. 67–71):

1. All citizens are owed an honest and publicly accessible justification for the use of political power in their society
2. So our conception of justice must be publicly accessible.
3. In order for our conception of justice to be publicly accessible, it must be justified by arguments that exclusively incorporate generally accepted beliefs and value-assertions.
4. So our conception of justice must be justified by arguments that exclusively incorporate generally accepted beliefs and value-assertions.

Wall's objection to this argument focuses on (3). He makes the useful distinction between arguments that are publicly accessible, publicly understandable, and publicly acceptable (Wall 1998, pp. 109–111). An argument is publicly accessible just in case it makes exclusive use of reasons and evidence that can be publicly stated and evaluated. An argument which appeals to the apprehensions of a particular person's private moral sense, for example, would not be publicly accessible. An argument is publicly understandable just in case it is publicly accessible and it does not include premises that the average member of the society is unable to understand.

[5] We might be tempted to think that Rawls could defend this claim by appealing to the priority of the right over the good. This move, however, is surely not available to the political liberalism of the later Rawls. The claim that political theory must prioritize the right over the good is a controversial philosophical one that would be excluded from the overlapping consensus.

Finally, an argument is publicly acceptable just in case it is publicly understandable and not subject to reasonable disagreement.

Wall observes that Rawls' argument does not go through if we only insist on publicly accessible arguments. The argument requires that we rely exclusively on publicly acceptable arguments. I think there is a strong case to be made for the claim that the members of a society are owed honest and publicly accessible arguments for the exercise of political power. But this does not conflict with perfectionism. To insist on publicly understandable arguments may already eliminate much good policy—the world is, after all, a very complicated place—and that may be too high a cost to justify in some cases. Why, then, should we go even further and insist on public acceptability, and thus accept argument-neutrality? Rawls' answer is that public acceptability is required to realize the value of "full political autonomy": a state of affairs in which every member of a society endorses and identifies with the social conception of justice (Rawls 1987, pp. 77–78).

Wall's objection to this argument parallels his objection to the argument from toleration. The argument seems to require the implausible claim that the value of full political autonomy will always outweigh the cost. But again, there is another possibility. One might argue that a commitment to argument-neutrality is a necessary constituent of a distinctively *liberal* political theory. Again, I defer addressing this possibility until Chap. 16.

3.2 Perfectionism and Liberty

Can a perfectionist theory of justice commit itself to the essential tenets of liberalism without thereby becoming incoherent? My goal in this section and the next is to take us a significant portion of the way toward answering this question in the affirmative. We will be left, however, with a number of objections whose resolution will have to wait until Chap. 16.

I begin with a very brief presentation of the thorough-going Aristotelianism of Martha Nussbaum's view of distributive justice, which she has named Aristotelian Social Democracy (Nussbaum 1988a, 1990, 1993). I then introduce David Charles' liberty-based objection to Nussbaum's view (Charles 1988), and argue that Nussbaum's response is inadequate (Nussbaum 1988b). I argue that though Nussbaum's political perfectionism cannot cope with the objection, the liberal is not thereby forced to embrace neutrality. A perfectionist theory that accepts the Principle of Competitive Value Pluralism, instead of the Principle of Neutrality, can avoid the objection. Finally, I consider Richard Arneson's argument that a commitment to value pluralism lends no support to liberalism, and argue that it does not apply to my formulation of value pluralism.

The two Aristotelian claims that lie at the heart of Nussbaum's theory of distributive justice are as follows:

1. A political arrangement of a society is good just in case, and to the extent that, it secures for each of the members of that society the necessary conditions of the good life (Nussbaum 1988a, pp. 146–147).
2. At some suitable level of generality, there is just one list of functionings that constitute the good life, and the contribution of any given functioning to the good life can be objectively assessed (Nussbaum 1988a, pp. 152, 176).

The characteristic feature of Nussbaum's view, then, is its *monism*; she rejects the idea that good human lives come in many forms, at least some of which are mutually exclusive.

Charles' liberty-based objection to Nussbaum's view focuses on the fact that her view endorses the claim that we can rank-order possible forms of life according to their level of excellence by assessing the extent to which each of the functionings that constitute those forms of life contributes to the good life. Charles asks "If the perfectionists [*sic*] considers some lives as better than others, can he avoid favouring coercion to force people into excellence?" (Charles 1988, p. 202). Although he recognizes "Aristotle's insistence…on choice at the final stage of development," he observes that "this seems in no way to lessen the possibility of unwarranted coercion at the earlier stages" (Charles 1988, pp. 202–203). Nussbaum's only response to this is to assert that Aristotelian habituation "need not" be coercive, and that it might in fact be essential to developing the capacity to make free choices (Nussbaum 1988b, pp. 212–213). Both these claims may very well be true. But the problem remains of what should be done, on Nussbaum's view, when someone does not, in the end, come around to freely choosing the good life.

Charles suggests that the only way to avoid an endorsement of coercion in cases where it is intuitively unwarranted is to go the route of the nineteenth century British perfectionists such as T.H. Green, who believed that "proper excellence can only be achieved by the free choices and actions of the people themselves" (Charles 1988, p. 203). It is not clear, however, that Nussbaum's view could justifiably integrate a commitment to free choice as a necessary constituent of the good life. The reason for this is her view's endorsement of value monism. In order for free choice to be part of a good life, it must be possible to freely choose to lead a good life. But freedom of choice requires access to a range of options. If there is a single form of life that is the good human life, as Nussbaum claims, then the free choice of a good life will require that various inferior forms of life remain accessible as well. More than this, the State that has as its goal creating the conditions that will best enable its subjects to lead good lives will have to actively preserve the accessibility of inferior forms of life, so that the possibility of freely choosing the good life is also thereby preserved. And if this is the cost of integrating the value of free choice into a monistic perfectionism, it is doubtful Nussbaum will be willing to pay it.

This cost can be avoided, however, without abandoning perfectionism, if we reject value monism. Consider a perfectionism that endorses the *Principle of Competitive Value Pluralism*:

> There are many equally good ways of life, which are incompatible insofar as leading one excludes leading others, and the values that structure some conflict with the values that structure others.

The competitive aspect of the principle ensures that there is no level of generality at which we can say that there is really only one good form of life, and thus lapse back into monism. To endorse value pluralism is to endorse the claim that from the perspective of a society as a whole, there are many different forms of life that are equally good. This is consistent with the claim that individual members of a society, from their differing positions, will come to different conclusions about what sort of life is most choiceworthy as the result of engaging in ends-deliberation. The perspective of the pluralistic society is that any form of life which could be most preferred by one agent in the society, given that agent's position and as a result of a well-executed course of ends-deliberation, is just as good as any form of life that could be most preferred by another agent in the society, again given his position and as the result of a well-executed course of ends-deliberation.[6] A perfectionism that endorses competitive value pluralism can commit itself to preserving access to a broad range of options for all members of a society without committing itself to preserving options which, from the perspective of the society as a whole, are not worth choosing.

There is, however, a deeper problem. Let us suppose that it is beyond our powers to eliminate all options that are not worth choosing without adversely affecting the ability of individuals to pursue choiceworthy options. Even if we endorse competitive value pluralism, access to some options not worthy of choice will remain. Let us further suppose that some individuals, despite the best efforts of the society, persist in choosing these options, but do so without causing harm to anyone else. Then Charles' concern about unwarranted coercion still stands: unless we assert that free choice is a *sine qua non* of a valuable life, how can we, from within a broadly perfectionist framework, oppose coercing those who are wasting their lives into pursuing excellence? And as Charles recognizes, the claim that a life that is not freely chosen is not valuable, no matter what achievements it contains, is highly implausible (Charles 1988, p. 203). So the liberal perfectionist must find some way to oppose coercion in this case without relying on an over-inflated conception of the value of freedom. This is precisely the task to which Chap. 16 is dedicated. If the liberal perfectionist cannot find a satisfactory way to do this, then there is reason to believe, as suggested in the last section, that a commitment to neutrality about conceptions of the good life is a requirement of liberalism after all.

Before proceeding to discuss a different threat to the conjunction of liberalism and perfectionism, I want to discuss briefly the relationship between liberalism and value pluralism. I have just suggested that if the anti-coercion argument of Chap. 16 succeeds, liberals will be able to make due with a commitment to competitive value pluralism and stop short of a commitment to neutrality. Richard Arneson, however, has argued that a commitment to value pluralism does not support liberalism over

[6] The social preference-order over forms of life in a pluralistic society, then, is not an Arrovian aggregate of individual preference-orders—in fact, it is entirely separate from them.

totalitarianism in the least. Arneson accepts William Galston's characterization of the thesis of value pluralism. According to Galston, among the set of potential ends and goals of human life, "there are multiple goods that differ qualitatively from one another and that cannot be rank-ordered" (Galston 2009, p. 95). Galston belies that the fact of value pluralism as he defines it entails that "Any public policy that relies on, promotes, or commands a single conception of human good or excellence for all individuals is on its face illegitimate" (Galston 2009, p. 96). Arneson argues, quite correctly, that nothing like Galston's conclusion follows from adherence to value pluralism understood in this way:

> If there are plural values and no ranking of them can be defended, then one cannot claim that in organizing society to maximize the single value *X* one is maximizing what is best. But equally no one can *object* to making *X* the politically privileged value on the ground that better outcomes would be obtained if we let a thousand flowers bloom, so values *A* through *W* would be achieved, the great flourishing of these values being more than adequate for the loss in achievement of *X* that would accompany the liberalization of society… Incommensurability entails that we lack a scale on which such measurements could be made. (Arneson 2009, p. 929)

The desired conclusion, however, is precisely what follows from adherence to value pluralism as I have defined it. The difference between my thesis and Galston's is that I deny, as Arneson's argument shows the liberal must, the claim that the potential ends and goals of human life are incomparable.[7] My version of pluralism, again, is that any form of life which could be most preferred by one agent in the society, given that agent's position and as a result of a well-executed course of ends-deliberation, is just as good from the perspective of society as a whole as any form of life that could most preferred by another agent in the society, again given his position and as the result of a well-executed course of ends-deliberation. And this entails the rejection of incomparability. On my view, then, we are perfectly entitled to reject the prioritization of one form of the good life, on the grounds that (a) whatever form we might prioritize, it will be neither better than nor incomparable to the many other forms we could leave space for; and (b) the chances of a great many individuals achieving a high level of functioning are much higher if there are many sets of equally valuable functionings from which to choose, rather than a single small set.

3.3 Perfectionism and Distribution

Contemporary liberalism is generally thought to be at least partly constituted by a commitment to some form of egalitarianism. There must be some way in which all members of a liberal society are treated equally. Charles' second objection to the claim that liberalism and perfectionism are compatible focuses on this issue of equality. Charles issues two challenges to the liberal perfectionist:

[7] This is clearly what Galston means when he uses the vexed term "incommensurable."

(i) [W]hat resources can [perfectionism] allocate to those who lack the capabilities for a fully functioning human life, or more radically for any part of that life?
(ii) What role can perfectionism give to the satisfaction of the basic needs of less well-endowed people? (Charles 1988, p. 204)

Charles' challenges are in need of clarification and refinement. First, let us distinguish between the basic functionings required for survival, and the higher functionings which, together with the basic functionings, constitute the good life. Then, let us distinguish between those who, according to a monistic account like Nussbaum's, are not capable of achieving every sort of functioning which is part of the good life, and those who are not capable of achieving any of these functionings beyond the basic functionings required for survival. The first challenge then has the following two parts: (a) Can the perfectionist justify allocating resources to those who lack the capabilities needed to achieve all the higher functionings that constitute the good life? And if so, (b) Can the perfectionist justify allocating resources to those who lack the capabilities needed to achieve any of the higher functionings that constitute the good life? The second challenge, which is simply a more radical version of the first, likewise has two parts: (a) Can the perfectionist justify satisfying even the basic needs of those who lack the capabilities to achieve all the higher functionings that constitute the good life? And if so, (b) Can the perfectionist justify satisfying even the basic needs of those who lack the capabilities to achieve any of the higher functionings that constitute the good life? Charles fears that the answer to these challenges must be 'No.' He argues that "The perfectionist must consider some lives (taken as a whole) as of greater value than others. If so, can he avoid directing most resources to those with the best abilities and most valuable inclinations (in conditions of scarcity)? It seems not" (Charles 1988, p. 205). Instead, the perfectionist must prefer the greatest amount of excellent activity "measured either in average in or quantity terms" (Charles 1988, p. 204).

Nussbaum responds to this charge by asserting that "[I]f having his people reach *eudaimonia* is…the legislator's goal, then, even if *eudaimonia* does admit of degrees (a very unclear issue), the legislator will direct his energies to bringing as many people as possible above the threshold, rather than to augmenting the *eudaimonia* of those who are already above" (Nussbaum 1988b, p. 214). This 'satisficing egalitarian perfectionism' is vulnerable to a number of objections, and has been criticized by Arneson in the course of his defense of prioritarianism (Arneson 2000). I will postpone discussion of these problems until I come to my examination of the prioritarian view. There I will show that the prioritarian view cannot be formulated in a way that makes it well-motivated, and so Arneson's arguments for what he wrongly takes to be a cogent view are superfluous. His criticisms of satisficing egalitarian perfectionism, however, are quite instructive, and will help us to identify a number of pitfalls to avoid.

For now, I conclude this section with my own response to Charles' challenges. Charles is right that the perfectionist must consider some lives better than others. Even given a commitment to competitive value pluralism, it will be the case that some forms of life are judged from the perspective of society as a whole as not

worthy of choice. But it does not follow from this, or from any other feature of perfectionism, that the perfectionist must favor the distribution that leads to the greatest amount of excellent activity, whether on average or in the aggregate. Charles' mistake is his failure to distinguish between *dividendum*—the object of the State's distributive concern[8]—and *modus dividendi*—the scheme according to which the State seeks to distribute that which it is concerned with distributing. Let us assume for the time being that the perfectionist is concerned with the distribution of achieved functioning. I argue below that this is not quite right, and that the perfectionist can, should, and must be concerned with the distribution of the liberty to achieve functioning. There is nothing about the nature of achieved functioning *qua dividendum* that makes it the case that maximizing average functioning or aggregate functioning is what the perfectionist must be committed to. This point is completely general. There is likewise nothing about pleasure, utility, primary goods, material resources, income, or any other possible object of distributive concern that implies a commitment to a maximum average or a maximum aggregate (or a maximum product, a maximum median, etc.) An advocate of any *dividendum* can perfectly coherently advocate, for example, a *modus dividendi* according to which the goal is the greatest attainable equal distribution.

The notion of a greatest attainable equal distribution is a familiar one in modern welfare economics, which takes utility—understood in terms of the satisfaction of ordinally ranked revealed preferences—as the object of distributive concern. The *Pareto frontier* is the set of attainable utility-distributions such that it is impossible to increase the utility of one agent without decreasing the utility of another—i.e. the set of Pareto efficient (or equivalently, Pareto optimal) distributions. On the graph below, the greatest equal distribution is the point on the Pareto frontier at which the utility of Agent 1 is equal to the utility of Agent 2 (Fig. 9.1).

Because the Pareto frontier in this case is concave, the greatest equal distribution maximizes neither average utility nor aggregate utility. But it remains a perfectly justifiable policy goal nonetheless for the advocate of utility as the appropriate object of distributive concern. What makes the perfectionist a perfectionist is the fact that he is concerned with valuable functioning—whether actual achievement of functioning or the liberty to achieve it. The assertion that the perfectionist must be an average or aggregate maximizer, and so cannot be an egalitarian, rests on a confusion. Even if perfectionist views are not inherently maximizing, however, the question of how to justify the policies Charles describes remains. But as we will see in Chap. 11, the Theory of Equal Liberty has no difficulty answering all of Charles' questions in the affirmative.

[8] The object of the State's distributive concern should of course not be confused with that which the State actually distributes. What the State actually distributes is *resources*—material capital, services and access to services, social and human capital, etc.—and legal rights. A maximizing hedonist, for example, is concerned with maximizing pleasure, and so searches for the distribution of resources that will result in the greatest attainable total amount of pleasure. He is not under the erroneous impression that pleasure itself can somehow be doled out directly.

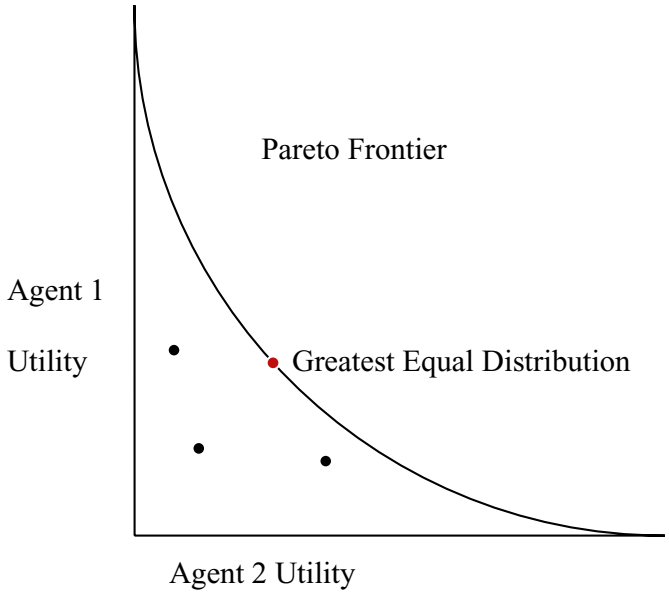

Fig. 9.1 Greatest equal distribution of welfare on a concave Pareto frontier

4 Utility, Priority and Equality

4.1 Utilitarianism and Exploitation

Utilitarianism, taken as a distributive theory, is both a theory of the best scheme of distribution (the *modus dividendi*) and a family of theories of what the object of our distributive concern should be (the *dividendum*). The distributive scheme is simple aggregate value maximization. The candidate objects of distributive concern are pleasure, happiness, or utility, with the last understood as either actual preference satisfaction, or "ideal" preference satisfaction. The last of these is often thought to make for the most plausible version of utilitarianism, with the ideal agent understood as one who is fully rational and informed. If we take the satisfaction of the preferences that agents would have if they were fully rational and informed to be the utilitarian's concern, then, his view has three essential parts: (a) that something satisfies a preference of a fully rational and informed agent makes it the case that it is good—and the higher the preference, or the more agents whose preferences it satisfies, the better it is; (b) the best social arrangement is the one that maximizes aggregate ideal preference satisfaction; and (c) preference satisfaction can be represented by cardinally measurable and interpersonally unit-comparable (or fully comparable) utility-values (this is needed in order to make sense of aggregate maximization).

Though I doubt it is often thought of in these terms, it seems to me that to endorse ideal preference theory—to endorse the claim that what makes something good is

the fact that it is preferred by a fully rational and informed agent—risks impalement on one of the horns of the Euthyphro dilemma. For suppose (1) that an omniscient god exists and (2) that this omniscient god has preferences over the ways that the lives of each of his creatures could go. This god is fully rational and informed, if anyone is. In fact, by defining ideal preferences in terms of the preferences of such a god, we eliminate any objections to the theory based on the limited cognitive capacities of human beings. So a very strong way to formulate ideal preference theory would be to say that, supposing such a god existed, the fact that something satisfied one of his preferences for one of his creations would make it the case that that thing is good. And the higher the preference satisfied, or the greater the number of preferences satisfied with respect to a greater number of his creations, the better it is. But this is exactly analogous to the first horn of the Euthyphro dilemma: what is pious is so because the gods love it, and the more they love it, the more pious it is. The contemporary ideal preference theorist is in just as bad a position as Euthyphro was to explain *why* the love of the gods *makes* anything good; and even more difficult, surely, would it be to explain why the love of *humans*, however rational and informed they may be, is capable of doing so.

This argument against utilitarianism as a theory of the object of distributive concern is far from decisive. As I have already noted in the introduction to this part of the book, I do not think there is any decisive argument against utilitarianism. My next argument, which concerns the mode of distribution rather than what is distributed, is more general. It is an argument against any simple aggregate value maximizing view. It is not decisive either; but making it allows me to specify one of the basic criteria which I take to be essential for any adequate theory of distributive justice. And even if this view is not universal, I believe it is widely shared.

John Roemer has given us the most plausible and precise account to date of what it means for one group of agents to *exploit* another (Roemer 1982, pp. 194–195):

> Suppose society is divided into the two groups *C* and *W*. Group *C* exploits group *W* if
>
> 1. There is an alternative social arrangement, which we may conceive of as hypothetically feasible, in which group *W* would be better off than in its present situation.
> 2. Under this alternative group *C* would be worse off than at present.
> 3. Group *C* is in a relationship of dominance to group *W*. This dominance allows it to prevent group *W* from realizing this alternative.

The relevant sort of alternative social arrangement is one in which the members of group *W* withdraw from society in accordance with a withdrawal rule. This is thus a game-theoretic theory of exploitation. In multi-person games that allow coalitions, we can define the "core" of the game as that set of solutions to the game in which no possible coalition of players could improve their situation by withdrawing from the game. The withdrawal rule concerns what the withdrawing agents may take with them when they withdraw—whether they may take only their labor; or their labor and their *per capita* share of the alienable means of production (transferable goods such as land and machinery); or these as well as, *per impossibile*, their *per capita* share of the inalienable means of production (the natural potential to develop abilities). Capitalism is supposed to abolish feudalistic exploitation by

allowing the individual to withdraw his labor from his employer, and go work for another or for himself as he wishes. Socialism is supposed to abolish capitalistic exploitation by eliminating differential ownership of the alienable means of production. And the dream of communism is supposed to be to abolish even socialist exploitation based on differential possession of the inalienable means of production ("from each according to his ability, to each according to his needs.")

It is not my view that the elimination of differential ownership of the alienable means of production—let alone the neutralization of differential possession of the inalienable means—is the only, or even the best, non-exploitative system of social organization. Chapter 11 will set out a group of distributive goals the achievement of which would realize a system of social organization that no members would wish to withdraw from, regardless of what they were allowed to take with them; and will argue that achieving these goals does not require abolishing differential ownership. For the present, however, all I wish to note is that a utilitarian society, or any society based on simple aggregate value maximization, is consistent with the existence of exploitation. These theories thus fail to be non-exploitative, which as I have already said, is one of my fundamental criteria for any adequate theory of distributive justice. The reasoning behind this claim is familiar enough. If the aggregate can be maximized by enslaving a small portion of the population, so that the rest are free to lead lives of great value without having to worry about how their own mundane needs will be met, then the aggregate maximizing approach recommends this arrangement. But such an arrangement clearly satisfies the definition of exploitation.

Even if it were never the case that maximum utility coincided with exploitation—even if, given some basic facts about human nature and the physical world, it were impossible—this would be no consolation to the utilitarian. For this would only show that a distribution that satisfies the utilitarian criterion *happens to coincide* with a non-exploitative distribution. If we accept the criterion of non-exploitation as a fundamental criterion of distributive justice, we cannot claim, as the utilitarian must, that a society is just *in virtue of* satisfying the utilitarian criterion. There is no getting around the fact that utilitarianism and exploitation are conceptually compatible. On the other hand, non-exploitation is an integral part of the egalitarian's approach to distributive justice. Any theory of distributive justice that does not minimize as far as is feasible the potential for exploitation cannot, for that very reason, be a satisfactory egalitarian theory.

4.2 Prioritarianism

Derek Parfit has proposed an approach to distributive justice, which he calls "the priority view," and which has come to be known as "prioritarianism," which seeks to provide an elegant way of eliminating the counter-intuitive demands of both aggregate value maximization and strict egalitarianism (Parfit 1997). Suppose we have a sound theory of well-being, of what makes a life good. Let us therefore

suppose for the moment that the question of the appropriate object of distributive concern—the *dividendum*—has been settled. We will focus exclusively on the question of what mode of distribution—the *modus dividendi*—to endorse. Our distributive goal could be to maximize aggregate value, the overall total amount of the *dividendum*. This approach is often thought ethically unappealing, for reasons like those just discussed. Or our distributive goal could be to achieve strict equality among all individuals. Such a goal, however, ignores, among other factors, the fact that in some cases strict equality can only be achieved by harming the better-off so that they are brought down to the same level as the worse-off. This is the so-called "leveling-down" objection to strict egalitarianism, which I will discuss in more detail later on. This approach is also often thought to be ethically unappealing.

Parfit's guiding idea is that *greater moral worth* attaches to aiding those who are in need than attaches to doing a good turn for those who are already reasonably well-off. Suppose we have a society made up of a set of individuals $I = 1, 2, \ldots n$. We are distributing resources among the members of this society, and we are interested in how good a life each individual will be able to lead given a particular bundle of resources. Let g_1 be the resource given in individual 1, and $v_1(g_1)$ be the value to individual 1, in terms of the goodness of the life they enable him to lead, of those resources. The aggregate maximizer defines an additively separable total-value function $V(g_1, g_2, \ldots g_n)$, and tells us to distribute—i.e., to choose $\boldsymbol{g} = (g_1, g_2, \ldots g_n)$—so as to maximize this function:

$$\max V\left(g_1, g_2, \ldots, g_n\right) = \max \sum_{i=1}^{n} v_i\left(g_i\right)$$

The priority view, however, defines a *moral worth coefficient*, w. The coefficient w_1 represents the moral worth of giving individual 1 a given bundle of resources. The prioritarian, then, gives us the following maximization problem to solve:

$$\max V\left(g_1, g_2, \ldots, g_n\right) = \max \sum_{i=1}^{n} w_i\left[v_i\left(g_i\right)\right]$$

And the prioritarian stipulates that the worse off someone is, the more moral worth attaches to aiding him—priority for the worse-off is built into the view. In the familiar example of the society with a large elite supported by a small slave population, the prioritarian's welfare function would almost certainly return a greater value if those who are enslaved were set free and a drastic scheme of downwards redistribution put into effect. This is because the moral worth of giving to those who have very little would be extremely high, whereas the moral dis-worth of taking from those who have a lot would be very low.

Most critiques of prioritarianism focus either on the social choice-theoretic assumptions needed to generate a representation theorem for the prioritarian welfare function, or on hypothetical cases in which the view yields outcomes which

many are likely to find unappealing.[9] But perhaps a more fundamental case to be made against it. It is unclear whether prioritarianism has any positive motivation. Unlike utilitarianism and egalitarianism, there does not seem to be any deep challenge to which it answers, any deep value commitment or view of the fundamental demand of justice which it expresses.

To see how difficult it is to motivate prioritarianism, the first thing we must do is note that the valuation functions, v_i, capture the full extent of the value of a given distribution for a given set of agents. The value of the moral worth coefficient, then, does *not* capture the fact that giving a certain amount of aid to someone who is badly-off does more good for that badly-off individual, in the sense that it makes a bigger difference to the value of the life he will be able to lead, than giving the same amount of aid to someone who is already well-off would do. The fact that the aid will mean more to the one who is badly-off is already taken into account by the individual valuation function. The distributive goal of the prioritarian is to maximize a universal valuation function V, which is equal to the sum of the individual valuation functions weighted by moral worth *as judged from the perspective of the universe*.

Prioritarianism needs some way to motivate its claims about moral worth. The whole point of the view is, after all, to capture the moral importance of prioritizing those who have less, even when we could do as much or more to enrich the lives of those who are already doing well. The first point the prioritarian would want to make is that it is important to prioritize those who have less *because they have less*, and thus, that what we do for them makes a bigger difference to the quality of their lives than what we could instead do, given the same resources, for those who are already well-off. It makes such a big difference in virtue of the fact that at present they have so little. But the prioritarian must ground the moral worth of prioritizing the worse-off in something other than that consideration. For what distinguishes this view from simple aggregate value maximizing theories is the fact that it must claim, in at least some cases, that we ought to prefer aiding one who is badly-off even if this means we neglect to make an even greater contribution to the well-being of one who is well-off. Egalitarians have an obvious way of grounding a commitment to aiding the worse-off. We prefer to help the worse-off because we *value equality*. We value equality because we recognize it as the fundamental demand of justice that we not tolerate a state of affairs in which some, through no choice or fault of their own, fair worse than others. But this is precisely what the prioritarian cannot assert, since his view does not recognize equality as a demand of justice. For the prioritarian, prioritizing the worse-off is offered as a replacement for valuing equality and working toward realizing it. This is clear from the fact that the view

[9] By the far the best work in this vein has been done by David McCarthy, who has developed precise distinctions between formulations or prioritarianism consistent with Parfit's initial ambiguous discussion, stated and proved a prioritarian representation theorem, and given the most enlightening discussion to date of the problematic nature of some of the assumptions required by the theorem (McCarthy 2006, 2008). But I hold out little hope for the potential of debates over the intuitive appeal of social choice-theoretic assumptions to arrive at any stable conclusion.

allows for great benefits to the well-off to outweigh small benefits to the badly-off even given the disparity in the moral worth attending benefits of these types. The prioritarian must, therefore, search for another ground for his claims about the moral worth of aiding the worse-off. He must claim that prioritizing, but without strictly preferring, the improvement of the position of the worst-off is itself a value worth endorsing—what we might call the *value of priority*—and that this value should replace both the *value of equality* and the value at the heart of simple aggregate maximizing theories—which we might call the *value of totality*—as our guiding distributive consideration.

Evidence for such a conclusion would have to consist of a convincing case in which (1) there is a genuine difference between the prioritarian distributive outcome and the distributive outcome of a sensible egalitarian view; and (2) it is clear, from a pre-theoretic perspective, that the prioritarian distribution is better—more just, more socially desirable, of greater moral worth, etc. Since almost all of the cases aimed at egalitarian theories fail to elicit a counter-intuitive conclusion from my theory of Equal Liberty—as I will have demonstrated by the end of Chap. 11—such a case is hard to come by. In fact, the only one I know of is one proposed by Richard Arneson, who converted to prioritarianism after a long period of advocating equality of opportunity for welfare (Arneson 2000, p. 58). But the background assumptions it requires are so restrictive as to render it toothless.

Consider a case of two agents. They are both doing well, by whatever measure we have chosen to use—possession of resources, access to opportunities, preference-satisfaction, achieved functioning, etc.—but one is doing better than the other (in whichever of these categories we are using, or in all of them if we like). The worse-off one is in possession of some good that is of *barely* any value to him—Arneson uses the example of a pouch of flower seeds. He might plant them, but would get very little enjoyment from the flowers they would grow into. The better-off one, however, would have a profound aesthetic experience from the flowers that could be grown from those seeds. Arneson suggests that if the *only* options are leaving the seeds with the first agent, or transferring them to the second, we ought to prefer the latter. The egalitarian distribution may not be morally objectionable, but the prioritarian one is nonetheless better.

That the seeds should be given to the better-off one may seem plausible at first. But Arneson is only able to fuel the intuition behind his conclusion by assuming the absence even of an exchange economy. If the better-off person had something he could give the worse-off one in exchange for the seeds which the worse-off would benefit from more than he does from possessing the seeds—and presumably he would have something, if he really is better-off—then they ought to, and ought to be allowed to, make that exchange. And this is precisely what a maximin egalitarianism (of which, as we will see in the next chapter, my theory is a version) would recommend. If Arneson has shown that prioritarianism is the superior view under that extraordinary assumption, then so be it. And it does not help Arneson's case if we alter the example so that the problem is for some third-party to decide which of the two individuals will be the recipient of the unclaimed seeds (though this change does alleviate concerns about property-rights which the original case might be vul-

nerable to). The right decision would be to give them to the worse-off one so that he would have something to exchange. Arneson has failed to give us a reason to favor prioritarianism, given the existence of any economic mechanisms other than direct transfer by the State of endowment resources. This case is of no practical importance. As Abba Lerner has argued, in a static, pure exchange economy, the principle of diminishing marginal utility implies that the utility-maximizing distribution is the strict egalitarian one (Lerner 1944). The same is true of the prioritarian distribution. It is only in the context of societies with dynamic production economies—the type we actually find in the modern world—that the distinction between these theories becomes interesting, and the choice between them becomes important.

Given the existence of a production economy and the fact of diminishing marginal utility, many cases in which we face a choice between great benefit to the well-off and slight benefit to the badly-off will be ones in which tremendous transaction costs attach to aiding the badly-off. For instance, in the event of a natural disaster, or in the aftermath of military aggression, there may be a part of the population that is cut off from the usual lines of communication. Simply reaching these individuals, so that aid can be delivered to them, may involve tremendous costs. The amount of resources available for aid at any given time will be fixed. If the transaction costs are high enough, then, the amount of aid that is actually delivered to those in need may be quite small. But no argument should be needed to assert that we should not then choose to give those resources to the better-off instead, even though they would likely benefit them more, precisely because they could be transferred to them directly with very little lost to transaction costs. The right decision is to balance the urgent need to deliver a subsistence level of aid and the need to begin immediate reinvestment in infrastructure, so that future aid can be delivered at lower cost. Over time then, the situation will be transformed into the normal one in which greater benefit can be done for the badly-off than for the well-off, given the resources available at any given time.

We can make Arneson's case more difficult—as well as give it more bite—if we imagine that what is to be distributed is an extremely scarce resource which we have good reason for wanting to keep out of private exchange—such as a medicine which can only be generated slowly and in small quantities. Suppose we have good reason to believe that the dose on hand will completely cure the chronic pain of a better-off individual, but only give slight relief to a worse-off one. We may assume that the difference in well-being is wholly down to a difference in the severity of the pain, or that there are other differences besides. This sort of healthcare ethics question might seem like the natural domain of prioritarianism. But this appearance results from an unjustifiable narrowness of focus. To give the dose to the better-off individual may well be the right decision; but the broader context in which this decision would be just is one in which the worse-off individual is compensated, through quality-of-life insurance funded by a tax-and-transfer mechanism, so that he is in the end no worse off (at least) than he would have been had he gained the (admittedly slight) relief of the medicine. This is precisely as maximin egalitarianism would recommend.

One might of course object that distributive problems sometimes involve non-transferable goods, and wonder whether Arneson's argument can be made to stick in such a case. Suppose, for instance, that the problem involved creating a non-transferable opportunity for one agent or the other, and that the only options were to create an opportunity for the better-off agent which he would value highly, or one for the worse-off which he would barely value. But this scenario is actually no more amenable to Arneson's argument than is the original. There is still the potential for an exchange. The question now simply becomes: What bundles of transferable goods would the better-off agent be willing to give up in exchange for having this opportunity created for him? Ideally, the bundle of such goods which would make the biggest difference to the worse-off individual is what would be distributed to him in exchange for his forgoing the creation of the opportunity he would not value greatly. Again, this is in exact accordance with a maximin egalitarianism.

In the absence of any compelling cases in which prioritarianism can be shown to solve problems inherent in a sensible egalitarian view, it is doubtful that there is any case to be made for it, unless some new argument is found for a value of priority which is more appealing as a fundamental distributive value than the values of equality or totality. And again, by the end of Chap. 11, I will have discussed the broad range of cases introduced in the work of Parfit and Arneson—prioritarianism's two most prominent and, in my estimation, creative proponents—and shown that my version of egalitarianism is able to handle them all satisfactorily. Let us proceed, then, to examine a few of the objections that have been made against two popular egalitarian theories—considered as theories of the *modus dividendi*—which do succeed against those theories. These are objections which my own theory will have to avoid.

4.3 Strict and Satisficing Egalitarianism

Prioritarians were prompted to attempt to formulate their view by the problems they identified in various types of egalitarianism. Since the view I will ultimately present is a species of egalitarianism, I would do well to highlight these problems; they are pitfalls which I must be careful to avoid. If there is no way to formulate an egalitarian theory so as to avoid these problems, then that failure may itself be a reason to adopt prioritarianism, at least provisionally, despite the problem of grounding its notion of moral worth. I will here discuss objections to two types of egalitarianism: strict egalitarianism, according to which we should always prefer an equal distribution of the object of our distributive concern to an unequal one; and satisficing egalitarianism, which we have already seen Martha Nussbaum advocate in a perfectionist context. I will not answer these objections here. I doubt that they can be answered by the views against which they are directed. I will show that they do not affect the view I endorse, but this must wait until I have completed my critical survey of rival views and presented my own view. The objections are, as I said, pitfalls to be

avoided; my purpose here is to register them and to use them to show why neither strict nor satisficing egalitarianism will do.

The objection to strict egalitarianism that motivated the priority view is the leveling-down objection (Parfit 1997). Consider a society with two groups of equal size, A and B. Now consider the following two possible distributions (of resources, functioning, utility, or what have you):

1. Each member of group A—1
   Each member of group B—1
2. Each member of group A—9
   Each member of group B—8

It seems that the strict egalitarian view would have us prefer distribution (1) to distribution (2). Distribution (1) is equal, while distribution (2) is not. But a theory of distributive justice that prefers (1) to (2) is hardly plausible. So any form of egalitarianism we endorse must repudiate the leveling-down method of attaining equality.

Arneson frames another important objection against strict egalitarianism. The objection has two parts: (a) there are cases in which it is best not to transfer resources from the better-off to the worse-off; and (b) there are cases in which it is best to transfer resources from the worse-off to the better-off. For a case of type (a), consider an island society in which one group is living barely above subsistence, and the other is starving to death. If the distribution of food were equalized, those who are starving to death would not receive enough to avoid starvation, but those who are currently subsisting would also starve. In this case, strict equality is certainly not morally appealing (Arneson 2000, p. 55). For a case of type (b), consider a group of badly-off individuals who are in possession of some crop seeds which cannot grow in the soil where these individuals live. The seeds, however, would sprout in the soil where another group of individuals, who are slightly better-off but still close to subsistence, live. The additional crop seeds would bring them comfortably above mere subsistence. Assume it is possible to transport the seeds but not possible to relocate the badly-off individuals. Arneson believes, I think correctly, that the best thing would be to give the seeds to the better-off, even though this would be a move away from strict equality.

Nussbaum's satisficing egalitarianism does not advocate leveling-down. On that view, the goal is to bring as many individuals as possible past a threshold level of capability. We are not justified in taking so much from those who are above the threshold that they fall below it, unless we thereby achieve a net gain in the number of individuals who are above the threshold. And once everyone is above the threshold, inequalities are no longer significant. Despite avoiding the leveling-down objection, satisficing egalitarianism has some weaknesses, and has been criticized by Arneson from a prioritarian perspective. He makes two significant objections. The first concerns a choice between helping those who are just under the threshold and helping those who are well beneath it. Suppose we could either bring one person above the threshold, or do even more to help someone who was well below it without quite bringing him up to it. On Nussbaum's view, we must help the first

person. Our goal is to bring as many individuals as possible past the threshold. But Arneson plausibly counters that if we can instead do as much or more for someone who is worse off, then that is what we ought to do (Arneson 2000, p. 56). Next, consider a choice between helping someone who is just past the threshold and someone who is well above it. On Nussbaum's view, at least as she has articulated it, these two options are equally good. Both have attained a satisfactory level of capability, and so there is no reason to favor one over the other. But again, it seems that we should prefer helping the worse-off (Arneson 2000, p. 57). In addition to the leveling-down objection, then, a plausible version of egalitarianism must deal satisfactorily with the cases proposed by Arneson.

We shall examine a number of other cases offered as challenges to egalitarian modes of distribution in Chap. 11. There, all cases, including the ones already raised, will be considered from the perspective of my own brand of egalitarianism. But we must remember that the task of articulating an egalitarian theory is not limited to that of finding a defensible egalitarian mode of distribution. Egalitarianism must also settle on an answer to the question of what the appropriate object of distributive concern is—the appropriate *dividendum*—as indeed prioritarianism and aggregate maximizing views must also do. It is to this debate that I turn in the remainder of this chapter.

5 Equality of What?

5.1 Well-being, Liberty and Desert

Any egalitarian theory of justice must answer the question of what we should seek to equalize. I will argue that we should take liberty as our *equalisandum*. For the present, we can define liberty-based egalitarianism as the view that the State should distribute resources so as to most effectively promote the equal development of autonomy, and equal freedom to exercise that autonomy by choosing, developing and exercising comparable sets of capabilities. In Chap. 11, I will discuss liberty-based egalitarianism in greater detail.

In addition to liberty—which is to say, the liberty to achieve well-being—we have the option of taking well-being itself, or the resources required to achieve well-being—that is, the desirable material conditions of life—as the appropriate *equalisandum* of an egalitarian theory of justice. The equality of liberty approach might be seen as standing between the other two, since the liberty to achieve well-being includes access to the required resources as well as the ability to transform those resources into functioning, the freedom to choose the functionings that will constitute one's way of life, and the capacity to deliberate well about which functionings to pursue. I begin my critical discussion with the equal well-being approach. Two possible forms of egalitarianism that I will not discuss are pleasure egalitarianism and preference-satisfaction egalitarianism. I do not think there is any good

argument for the former, and I have already argued against preference-satisfaction as the appropriate object of the State's distributive concern.

There is one argument against equality of well-being itself which we should be careful not to make, since it rests on a confusion. There is a sense in which each person's achievement must be his own (though this does not ground an argument for the absolute value of freedom, as Green thought it did) (Green 1883).[10] If an achievement is to be mine, then there must be some point at which I set my own limbs and my own mind in motion—whether I do so for fear of sanction or for any other reason is beside the point—in order to attain that achievement. If I gain rewards by taking credit for the work of others, or if I am physically compelled to perform some task in the sense that another overpowers me and moves my limbs for me, then I have achieved nothing. But it does not follow from the fact that each person's achievement of well-being must be his own in this sense that a society cannot be concerned with the distribution of well-being itself, but only with the freedom to achieve well-being. What the State actually distributes are resources and legal rights to possess or make use of them, where by resources we include not only material resources but also access to services and human and social capital. The reason for this is simple: resources are what we can actually move around, and legal rights exist in virtue of being conferred. But the State may very well take as its distributive goal equality of achieved well-being, and seek to distribute resources, broadly construed, so as to maximize the likelihood of this distributive goal being reached. The fact that each person's achievement of well-being must ultimately be his own does not show this goal to be an incoherent one.

The problem with the equality of well-being approach is that it does not take into account the issue of desert. If different individuals expend different levels of effort, and those individuals are responsible for the levels of effort they expend, then it is not just for every individual to be enabled to achieve the same level of well-being. A sophisticated egalitarianism must be sensitive to this concern. The appropriate way of accommodating this point from an egalitarian perspective may be expressed thus:

> The only thing about people's labor that would validate the justice of a difference in the level of well-being they get from it is a difference in the burden of that labor, broadly construed.[11]

The central idea of a sophisticated egalitarianism is that how well one's life goes should depend on how much effort one chooses to put into one's life, and not on features of oneself or one's environment that are "morally arbitrary," such as the fact

[10] More discussion of the value of freedom and the liberal argument against coercion will come in Chap. 16 *infra*.

[11] This is a paraphrase of GA Cohen, with "level of well-being" substituted for "income." After a time as an advocate of equality of well-being/functioning, Cohen returned at the end of his life to favoring equality of resources. But what is significant, for him, about having (or having access to) resources is the fact that this is required for leading a worthwhile life. For the original quote, see (Cohen 2008, p. 181). For Cohen on equality of functioning, see (Cohen 1993).

that one was born into a wealthy family or an excellent school district.[12] The simple equality of well-being approach, then, is not really a serious contender. But one might argue that the appropriate object of the State's distributive concern is well-being *subject to desert*. How does this view compare to one that advocates equality of liberty to achieve well-being?

One might be tempted to argue that, at the end of the day, a desert-based egalitarianism collapses into a liberty-based egalitarianism. Such an argument would run thus. It ought to be that those who deserve to achieve equal levels of well-being do achieve equal levels of well-being. We deserve what we work for, which is to say, how much we deserve in life depends on how much effort we expend, provided that we are responsible for our own level of effort. But we are responsible for our level of effort when we freely and autonomously choose what goals to devote ourselves to and how much effort to put into pursuing our goals. So the way to promote equality of well-being subject to desert just is to promote the conditions of equality of liberty, in which every individual has an equal degree of freedom to autonomously select and pursue his own goals. If this argument goes through, and we do in fact promote equal well-being subject to desert *via* promoting equal liberty, then we might still want a reason to consider the view under the guise of liberty-based egalitarianism rather than desert-based egalitarianism. And an excellent reason is ready to hand: namely, the fact that we now have at our disposal a rigorous, precise, and measurable conception of individual liberty. We gain a great deal of clarity by understanding deserved achievement as achievement attained from an initial position of liberty to achieve which is equal to the position of everyone else.

But not everyone will find this argument convincing. Some will insist that a focus on the distribution of liberty is inessential to a theory concerned with well-being subject to desert.[13] And I am inclined to think that the argument is at least a bit too quick. It may very well be that a particular individual is doing exactly as well in life as he deserves to be doing, based on his level of effort for which he is responsible, even though he lives in a society in which the liberty to achieve well-being is not equally distributed. It may even be that every individual in that society is doing as well as he or she deserves to be doing. But then we have a very good reason to doubt that equal well-being subject to desert suffices for distributive justice. The

[12] For an excellent discussion of the morally arbitrary, see (Cohen 2008, pp. 151–180). There are some liberals, such as William Galston and Brian Barry, who accept as legitimate differences in well-being due to morally arbitrary features. Galston asserts that morally arbitrary features of ourselves are often a large part of our own self-constructed identities (Galston 1991, pp. 196–197). This is true, but it is no reason to affirm the justice of disparities in well-being that derive from those features. Barry argues that our very existence is morally arbitrary and contingent, and so whatever level of well-being we achieve is due in part to an arbitrary feature (Barry 1988, p. 41). Barry's point, while correct, is of no practical importance. We are not faced with a choice between advancing the welfare of the existent or of the non-existent. We are faced with a choice between preserving the well-being of the well-off and advancing the well-being of the badly-off. Any arbitrary feature that is universally shared and shared to the same extent, like existence, need not be a source of concern for a theory of distributive justice. But that leaves plenty of morally arbitrary features that *are*.

[13] See, for example, (Arneson 1998, p. 190).

level of effort which it is reasonable to put into one's life depends on the goodness of the life one could realistically hope to lead, and on the quality of one's chances for leading that sort of life. If one's initial liberty to achieve well-being is limited compared to that of other members of one's society, it may simply be unreasonable for one to expend the same level of effort as they do, since one's own chances of leading the sorts of lives that are open to them are so slim.[14] So if desert and equal liberty really do come apart, the desert-based theory will always leave us with a troubling counter-factual. If a given individual's liberty to achieve well-being had been greater or less than it was, owing either to distribution up or distribution down for the sake of realizing equality of liberty, would he have achieved as much or as little as he has, and would he still deserve as much or as little as he has, in virtue of the level of effort that he would have expended in the counter-factual scenario? The danger of a desert-based theory that is divorced from a central concern for the distribution of liberty is that only for those who begin with a morally arbitrary head-start will it be reasonable to expend a high level of effort, and thereby deserve by the lights of the theory the high level of well-being which they achieve. Such a theory is forced to speak in terms of desert while turning a blind eye to the morally arbitrary differences in the positions from which individuals begin to strive for a good life. To speak of desert without concern for these differences is, at best, perverse, since the disparate starting points are not themselves deserved, and are likely to have a large impact on the levels of well-being that are ultimately achieved. If we are really concerned with well-being subject to desert, we ought to be concerned with desert all the way down. And this means being concerned with creating the conditions of equal liberty.

I have been arguing that if we believe that individuals should be able to lead lives as good as they deserve to lead, then the focus of our distributive concern should be on creating the conditions of equal liberty. But what one deserves is a function of the effort that one has expended for which one is responsible. Does making this argument, then, require affirming a connection between liberty and responsibility? And if so, does the argument run afoul of Frankfurt-style objections?[15] I think not. In Frankfurt's famous example, an individual faces a choice between two options. He has received a neural implant, unbeknownst to him, which will cause him to choose the option preferred by the mad scientist who inserted the implant should he be on the verge of choosing the dispreferred option. If he chooses the preferred option on his own, the implant will remain inactive (Frankfurt 1988). The point of the example is to show that one can be responsible for one's choices even if one does not have access to any alternatives—even if one lacks freedom of choice. The individual is supposed to be responsible for the choice of the scientist's preferred option in the case in which the implant remains inactive. But if this is so, it is presumably because that choice is autonomous. If the choice were the result of prefer-

[14] The issue of reasonable levels of effort will come up again in the discussion of John Roemer's notion of equality of opportunity for welfare.

[15] For a concern of this kind, see (Lippert-Rasmussen 1999).

ences that were coercively, oppressively, or manipulatively instilled in the agent, the issue of his responsibility for the choice would at least be considerably less clear.

The Frankfurt argument does not give us reason to doubt the relevance of autonomy to responsibility, even if it does give us reason to doubt the relevance of freedom of choice to responsibility. But even with respect to the latter connection, the force of the argument is limited. For in order for an agent to become autonomous, he must have the freedom required to develop that autonomy. And the development of autonomy, as discussed in Chap. 7, does require access to a broad range of options that enable the agent to perform a variety of experiments in living. And broad freedom of choice is required not only for the agent in question, but also for the other members of his society, whose testimonies about their own experiments in living are invaluable input to the agent's own process of autonomy-development. Broad freedom of choice, then, is required for one to become the kind of person who is capable of being responsible for his choices. And if the extent to which one deserves one's lot in life is a function of the extent to which one is responsible for one's choices and the level of effort one chooses to expend, then desert presupposes liberty.

Questions about the relationship between responsibility, freedom of choice, and autonomy inevitably lead to questions about the (in)compatibility of individual freedom and physical determinism. It is fair to ask whether the account of individual liberty I have developed, and the notion of responsibility I will make use of in filling out the equal liberty approach to distributive justice, are compatible with determinism. I believe that they are, but postpone this discussion until Sect. 5.4 below, where it arises in connection with Roemer's work.

5.2 Dworkin's Resource-Based Egalitarianism

I have been arguing against the idea that we can formulate a plausible theory of distributive justice that focuses on equalizing well-being subject to desert but that does not collapse into a theory that focuses on equalizing liberty. The consideration that drove us to consider equality of well-being subject to desert, rather than simple equality of well-being, was the fact that there are cases (or at least, it is possible for there to be cases) in which an individual is responsible for the fact that his life is not going as well as another person's. The notion of desert was introduced into the well-being equalizing view in an attempt to make room for individual responsibility. But the inadequacy of well-being subject to desert as a focus of distributive concern does not necessarily show that we must focus on the distribution of liberty to achieve well-being in order to accommodate the importance of individual responsibility. Ronald Dworkin's resource-based egalitarianism is designed specifically to deal with the problem of incorporating individual responsibility into a broadly egalitarian framework (Dworkin 1981). After briefly outlining Dworkin's view, however, I will argue that it is not, in fact, an alternative to liberty-based egalitarianism. Rather, it is a bridge between a Senian view of distributive justice as equality of

capability-freedom, and the fully developed view of equality of liberty which I will set out and defend in Chap. 11.

In discussing the idea of well-being subject to desert, I spoke of desert as a function of the amount of effort a person chose to put into his life, and began to argue that we cannot fairly say that a person has gotten what they deserve in life relative to someone else unless both individuals chose which pursuits to dedicate themselves to and how much effort to put into those pursuits from a position of equal liberty, using the rough characterization of the conditions of equal liberty given at the beginning of Sect. 5. The suggestion was that desert is a matter of the effort for which an individual is responsible, and responsibility for one's choices, at least in the very robust sense that a theory of distributive justice should be concerned with, requires liberty. Without conditions of equal liberty, those who start out already behind will always be able to complain, with justification, about the fact that they had to make their choices under a handicap, and so are at least not solely responsible for the outcomes.

This view of desert, however, is incomplete, and one of the merits of Dworkin's view is to identify another important component is judging desert and responsibility. Desert is not simply a function of chosen effort. One of the things we may choose to do in life is to take risks. If a risk is accepted freely and knowingly by someone with access to all the available information relevant to deciding whether to take the risk, and the outcome turns out to be a bad one, then no amount of effort on the part of the agent will make it the case that a better outcome was deserved.[16] Dworkin draws a distinction between what he calls "brute luck" and "option luck." Option luck concerns the outcomes of risks accepted freely, knowingly, and with access to all available information. Brute luck concerns the outcomes of actions and events under uncertainty in cases where the risk was not so accepted. One of Dworkin's central ideas is that it is just for individuals to be compensated for the negative outcomes of cases of brute luck, but not for the negative outcomes of cases of option luck. We are responsible for the risks we accept, as well as for the amount of effort we choose to put into our lives.

Dworkin's task, then, is to construct a scheme of distributive justice in which all individuals are treated equally, are held equally responsible for their own option luck and the amount of effort they choose to put into their lives, but are not held responsible for their brute luck. Dworkin begins by envisioning the members of a society placed behind a "thin" veil of ignorance—the sense in which the veil is thin will become clear below. He distinguishes between the circumstances of each agent, on the one hand, and his preferences and ambitions on the other. An agent's circumstances consist in his bundle of comprehensive resources, which include his talents and genetic advantages (or handicaps and disadvantages). Behind the veil of ignorance, each agent knows his own actual preferences and knows the overall distribution of various talents, advantages, handicaps and disadvantages in his society (hence the thinness of the veil) but does not know what talents/handicaps he has.

[16] Or, perhaps, in a situation in which any reasonable person should be expected to know that he is taking the risk.

Each individual is given an equal amount of currency (which Dworkin refers to as "clamshells"). They may then use this currency for two purposes. First, to purchase external, transferable resources which will help them to lead the sort of life they deem valuable, based on their preferences (which, again, are known to them). And second, to purchase insurance against being handicapped or disadvantaged—though not against lacking a talent or being unable to satisfy a preference. The idea here is that an agent will accept a conditional insurance contract only if the contract specifies that, should he find himself with a handicap or disadvantage when he steps out from behind the veil, he will receive some bundle of external, transferable resources which compensates him adequately for his bad brute luck. When the members of the society emerge from behind the veil, their purchases and conditional insurance contracts, made from behind the veil with equal initial monetary allotments, will induce a distribution of resources in the actual world. This, Dworkin claims, is the just distribution—one in which individuals are compensated for disadvantages due to their circumstances, but are held responsible for their choices.

One problem with Dworkin's view concerns his treatment of preferences and ambitions. Behind the thin veil, what purchases an agent makes and what insurance contracts he accepts are determined by his preferences (including his preferences over levels of risk). Dworkin must thus see preferences as something for which the agents can be fairly held responsible. He claims that an agent is responsible for one of his preferences so long as that preference is "authentic"—so long as the agent has had an ample and equal opportunity to form, reflect on, and defend that preference (Dworkin 1987, p. 35). This opportunity, Dworkin points out, requires that one possess an equal and extensive share of civil liberties—freedom from legal constraint on one's speech, conscience, etc (Dworkin 1987, p. 2). This is a good start, and a conviction shared by my own liberty-based egalitarianism. But it is only a start—a necessary, but quite far from sufficient, condition. If one is to be fairly held responsible for one's preferences, one needs much more than legal protection for one's civil liberties. One is responsible for one's preferences, in the way that matters for distributive justice, just in case those preferences are formed and adopted *autonomously*—and, as I will argue below, in precisely the way that I have explicated the notion of autonomy. This point is of significant consequence for the structure of Dworkin's view. Dworkin's social contract story assumes that the requirement that individuals' preferences be authentic can be satisfied prior to addressing the question of the distribution of tangible resources, by specifying the necessary legal framework of rights and freedoms prior to determining the answer to the distributional question. But as we will see in Chap. 11, the distribution of tangible resources, as well as legal rights and freedoms, is an integral part of the conditions of autonomous preference formation; and it is these conditions which must be realized if individuals are to be fairly held responsible for their preferences. So the resolution of at least some distributional issues must precede the negotiation of a social contract like the one Dworkin has in mind. Moreover, as we will likewise see in Chap. 11, we do not need to interpose a social contract mechanism at all in order to solve the problem of determining a just distribution of resources among individuals who are responsible for their preferences.

A second problem with Dworkin's view reveals the extent of its similarity to a Senian equality of capabilities. In fact, as we will now see, Dworkin's account of equality of resources just is such a view, augmented by the incomplete account of responsibility just discussed (which is why it can be seen as a bridge from Sen's work to my own.) Let us just assume that there is some way to incorporate genuine autonomy in preference formation and choice into Dworkin's account, and focus on his central proposal: that a just distribution is one in which individuals who are responsible for their own preferences are fairly compensated for their bad brute luck through the transfer of external resources. What question must an autonomous individual behind Dworkin's thin veil, who is contemplating whether a given insurance policy would justly compensate him for a given incident of bad brute luck, ask himself in order reach a sound conclusion? In order for the compensation received by the disadvantaged individual to be (rightly judged) just, he must receive a bundle of resources that will enable him to have the same opportunity to lead a valuable life, judged from his position, as he would have had if he had not been the victim of that bad brute luck. Whether or not the policy under consideration would provide compensation meeting this criterion in the relevant question.

But if that is right, then the judgment of whether compensation for bad brute luck is just or not depends on the nature and extent of the capability set which the agent has the opportunity to develop as a result of the resource transfer. The only way to fill out what it means for such a transfer to enable an individual to have the same chance at leading a valuable life as he otherwise would have had, is by specifying that he receive resources sufficient for a set of capabilities for functioning equal in extent to the ones enjoyed by those individuals of similar autonomous preference who have not experienced that bad brute luck. So long as we are concerned not merely with what people have, but with whether they are in a position to take what they have and use it in the course of leading a worthwhile life, the ultimate object of our distributive concern is capability-freedom rather than resources. Moreover, we have seen in Chap. 7 that it is possible to specify very precisely what it means to say that any two individuals possess (equal opportunity for developing) equal capability sets. That account is not dependent on any prior account of equality of resources with which to translate those capabilities into functionings, but it does allow us to give a wholly derivative account of equality of resources. Two agents are equal in their share of resources just in case the resources they possess provide them with equal opportunity sets of capabilities to develop, and subsequently, equal sets of capabilities to exercise. We cannot, however, reverse the order of explanation; there is no way to specify what it means to say that two individuals have equal shares of resources (in Dworkin's sense) without first determining whether they have equal shares of capability-freedom.

Dworkin's view thus turns out to be a species of liberty-based egalitarianism, rather than an alternative to it. At one point, Dworkin himself comes close to admitting as much: he maintains that Sen's capability-based egalitarian theory might not be a rival to his own:

> [A]ny differences in the degree to which people are not equally capable of realizing happiness and the other "complex" achievements should be attributable to differences in their

> choices and personality [for which they can fairly be held responsible] and the choices and personality of other people, not to differences in the personal and impersonal resources they command. If we do understand equality of capabilities in that way, it is not an alternative to equality of resources but only the same ideal set out in different vocabulary…But (on this reading of Sen's position) it is their personal and impersonal resources, not the happiness or well-being they can achieve through their choices, that are matters of egalitarian concern… the equality we seek is in personal and impersonal resources themselves, not in people's capacities to achieve welfare or well-being with those resources. (Dworkin 2000, p. 303)

The point Dworkin is trying to argue for is that the egalitarian only advocates equality of capability in virtue of advocating equality of resources; achieving the latter results in achieving the former, but the latter is the actual goal. As we have seen, it is Dworkin who gets the order of explanation wrong. The distribution of capability is the egalitarian's proper concern (given that individuals are held fairly responsible for their choices about which capabilities to develop and exercise). Equality of resources cannot play the role Dworkin has cast it in, since the very notion of equality in the distribution of resources derives its meaning entirely from an independently specifiable equality of capability. Dworkin's work does make some progress in spelling out what it means for individuals to be responsible for their preferences and choices, and this is an important contribution. But he fails to adequately represent the conditions required for the development of autonomous preference formation and choice, and so fails to develop a distributive mechanism which is appropriately sensitive to the distinction between the outcomes of autonomous choice and the effects of brute luck. An important step toward doing this has been made by John Roemer, whose theory we will examine in Sect. 5.4 below; and a fully adequate account is given by the theory I will develop in Chap. 11. But our next order of business is to try to look with fresh eyes at the theory developed by John Rawls, which, as we will see, can be interpreted as having been designed to cope with the possibility that the effects of autonomous choice and brute luck *cannot* be distinguished.

5.3 Rawlsian Equality of Liberty and Opportunity

As I discussed in the introduction, I have little to offer in the way of original criticism of Rawls' principles of justice. All I wish to do in this section is briefly register my agreement with complementary criticisms of Rawls' Principle of Fair Equality of Opportunity made by Arneson and Cohen, and then say something about the problematic nature of the Difference Principle from the perspective of intergenerational justice.

Rawls' Principle of Fair Equality of Opportunity concerns access to offices and employment. It states that individuals of equal talent and ambition should all have an equal chance to vie for such positions (Rawls 2001, p. 44). Those who are talented and ambitious among the economically disadvantaged should have access to

the same opportunities for education and training as the better-off. Among a pool of equally talented and ambitious applicants for a given position, placement should be determined solely by the particular requirements of the position with respect to skills and experience. My central question is: does the Principle of Fair Equality of Opportunity succeed in capturing the sort of equality of opportunity that we would expect to find in a just society? I think not. As Arneson has pointed out, this principle is beset by "the problem of stunted ambition" (Arneson 1999, p. 78). The problem is that Rawls' principle takes talent and ambition as given, and is totally insensitive to the fact that the potential for talent must be actualized and ambition must be developed and nurtured. As Arneson vividly explains:

> Fair Equality of Opportunity …is compatible with a further, disturbing description [of society]: all individuals are socialized to accept an ideology which teaches that it is inappropriate, unladylike, for women to aspire to many position of advantage, which are *de facto* reserved for men, since only men come to aspire to them. (Arneson 1999, p. 78)

The nurturing of talent and ambition is a process which is sensitive to social conditions. And as G.A. Cohen has forcefully argued, these social conditions extend well beyond the basic structure of society, to which Rawls' principles are limited (Cohen 1997). The Principle of Fair Equality of Opportunity fails to capture the conditions of equality in the distribution of opportunities for educational and professional advancement. Those born into cultural environments that nurture talent and foster ambition have an advantage in gaining access to these opportunities, and it is an advantage that is based on features of their lives that are morally arbitrary and beyond their control. And once again, we are moved toward the conclusion that what justice requires is that we focus on creating the conditions of equal opportunity to develop and exercise one's autonomy—and in particular, to make autonomous choices about which of one's talents to develop and how much effort to dedicate to developing them. These are precisely the conditions which are lacking in the scenario envisioned by Arneson.

We can extrapolate from Rawls' text the response that he would give to this criticism. Rawls did not think it possible to isolate the autonomous component of the effort an individual put into developing his abilities from the component that was due to the environment in which he was raised and his genetic endowment: "the superior character that enables us to make the effort to cultivate our abilities… depends in good part upon fortunate family and social circumstances in early life for which we can claim no credit" (Rawls 1971, p. 89). In the absence of any possibility of drawing this distinction, the only way to guarantee that differences in social position are due solely to differences in autonomous effort would be (*per impossibile*) to equalize the genetic and environmental positions from which all agents begin their lives. If we were to accept this conclusion, we would likely endorse the Principle of Fair Equality of Opportunity, despite its apparent shortcomings, as the best that we could do, and then seek to compensate for the residual unfairness that it leaves. That is one plausible interpretation of the motivation behind adopting the Difference Principle. The Difference Principle attempts to compensate through downward redistribution of wealth for the fact that some individuals are given more

support and encouragement than others to expend the effort needed to develop their talents. This interpretation of the Difference Principle as a compensatory principle has recently been emphasized by Dean Machin (Machin 2013). This explains why the Difference Principle does not instruct us to redistribute wealth in a way that is at all sensitive to desert: sensitivity to desert presumes the ability to distinguish an individual's level of autonomous effort. If, however, we can make these distinctions, then the Difference Principle's insensitivity to desert becomes a mark against it. In order to make these criticisms stick, then, we need a way of determining an individual agent's relative level of autonomous effort. One of the great advantages of John Roemer's theory of justice as equality of opportunity for welfare, the topic of the next section, is that is gives us a way of doing this.

If we can do better than the Principle of Fair Equality of Opportunity, then the argument for the Difference Principle, which is in part motivated by its supposed ability to compensate for the other principle's shortcomings, is also weakened. But there are other reasons to doubt it. Let us assume that the first two principles are already being observed. The Difference Principle then tells us that economic activity in a just society should be organized so as to maximize the worst-off person's share of primary goods. Primary goods include rights and liberties, health, income and the social bases of self-respect. But as a matter of fact, the Difference Principle will be concerned almost exclusively with income. The recognition of one's rights and liberties is secured by the first principle. Such recognition is one of the most important social bases of self-respect. The opportunity to seek meaningful work is another, but this falls within the province of the Principle of Fair Equality of Opportunity. The other important bases of self-respect, such as the ability to appear in public without shame, are best understood in terms of the power to purchase particular commodities. This is a point first observed by Adam Smith, in his discussion of the material needs of English as opposed to Scottish and French peasants:

> By necessaries I understand not only the commodities which are indispensably necessary for the support of life, but whatever the custom of the country renders it indecent for creditable people, even of the lowest order, to be without…Custom…has rendered leather shoes a necessary of life in England. The poorest creditable person of either sex would be ashamed to appear in public without them. In Scotland, custom has rendered them a necessary of life to the lowest order of men; but not to the same order of women, who may, without any discredit, walk about barefooted. In France they are necessaries neither to men nor to women, the lowest rank of both sexes appearing there publicly, without any discredit, sometimes in wooden shoes, and sometimes barefooted. (Smith 1776/2003, 5.II.2.IV)

If we interpret health, considered as a primary good, as access to quality healthcare, then this will either be a function of income—needed to purchase care—or of the State's level of tax revenue—needed to fund an adequate system of healthcare. If we take the primary goods with which the Difference Principle is concerned to be income and access to basic services, then, it seems that there are two very general possibilities for the economic implementation of the principle within a market economy. The first, and more obvious, is to set the marginal tax rate which is imposed on the highest earners to the rate which is most likely to maximize government tax revenue. This would give the Rawlsian government the greatest capacity for down-

ward redistribution of wealth and for the maintenance of basic services *within a given generation*. It is likely, however, that such a policy will reduce total government tax revenues over the long run, since it is likely that the top marginal tax rate that will maximize government tax revenue within a given year is somewhat higher than the rate that is most conducive to maximizing GDP.[17] Choosing that rate is the second policy option. Given that the agents in the original position are not assumed to be altruistic, it is unlikely that the second policy would be adopted rather than the first. This makes the implementation of the Difference Principle problematic. It is supposed to capture concern for the absolute position of the worst-off. But if the absolute position of the worst-off is maximized in the first generation (or even in the first year), the highest attainable share of the worst-off members of society in the long-run will be diminished. Yet this diminishment cannot be avoided without failing to maximize the share of primary goods of the worst-off members of society in the first generation, which, it seems, the Difference Principle instructs us to do, at least on the interpretation of it that is most likely to emerge from the original position. Nor does choosing the first policy seem to conflict with Rawls' intergenerational Just Savings Principle (Rawls 1971, §44). We will examine this principle, and the problem of intergenerational justice generally, in Chap. 11.

This problem aside, however, there is a deeper and simpler criticism of the Difference Principle. The principle, as we have seen, is primarily concerned with maximizing the income of the worst-off member of society. There is a substantial body of empirical evidence that speaks against this as a sensible policy goal. Past a low threshold, there is a very weak correlation between increases in income and improvements in physical and emotional health, subjective satisfaction and contentment, and other important components of overall well-being.[18] This is precisely the consideration that drove Sen to shift the focus of distributive justice from resources to capabilities—the valuable forms of life that an individual has the ability to realize given the resources he has access to. Rawls' principles, then, fail both to capture the conditions of genuine equality of opportunity, and to provide a sound guide for the distribution of resources and wealth.

5.4 *A Partial Defense of Roemer's "Equality of Opportunity for Welfare"*

John Roemer has developed a robust and sophisticated approach to resource distribution that takes the opportunity to lead a valuable life as the appropriate *equalisandum* (Roemer 1994). I will begin by briefly outlining his theory. I will then focus on the conception of individual responsibility that Roemer makes use of, and the way in which his model seeks to capture the role that individual autonomous effort plays in

[17] This is a complex issue, to which I will return in Chap. 11.

[18] For a classic study, see (Scitovsky 1976). For a more recent discussion with a wealth of econometric data and solid philosophical synthesis, see (Kenny and Kenny 2006).

determining the level of well-being an individual manages to achieve. I will argue, *contra* Mathias Risse, that Roemer can consistently adhere to a plausible compatibilist conception of responsibility. But I will then argue that Roemer's view fails as an *aspirational* account of justice. That Roemer takes himself to be constructing an aspirational account may be inferred from the fact that he sets his theory up as a competitor to the theories of justice advanced by Rawls, Sen, Cohen, and other aspirational theorists. Considered as an aspirational account, Roemer's view is problematic. But the possibility remains that the view is well suited to play the role of a transitional account; and I will argue at the end of this chapter that Roemer's view, considered as a transitional account, will prove to be an excellent match for my own aspirational account.

Roemer's view is quite complex, and I will only give a fairly brief and somewhat simplified overview of it.[19] The key idea is that individuals all find themselves in the midst of circumstances beyond their control, and the nature of these circumstances affects the amount of effort that they choose to put into making their own lives go well. He suggests that we divide society into *types*, based on external features that are outside the control of individuals, and which we recognize as being relevant to an individual's responsibility for the amount of effort that he or she chooses to put into pursuing his or her goals. In a society in which men and women are not viewed as equally capable, for example, gender is one such relevant feature. Other may include ethnic background, parents' annual income, etc. Let $I = 1,2,3\ldots$ be the set of types into which society is divided. Within each type, there will be a distribution of effort. Some individuals within a given type will put more effort into their lives than others. We may then speak in terms of the percentile of effort that a given agent falls in within his type. Roemer suggests that when we have succeeded in identifying all the external factors that are relevant to a fair judgment of an individual's personal responsibility for his own level of effort, we will see identical distributions of effort across types. If we see a higher average or median effort in one type than in another, we should take this as a sign that we have not yet captured all of the relevant external factors, and thus that we are not yet in a position to judge fairly individuals' personal responsibility for their efforts.

Let $\Pi = [0,1]$ be the set of effort-percentiles in any given type. Let ϕ be a scheme of distributing resources. And let $v(\cdot)$ be a valuation function, measuring level of well-being achieved. Roemer's claim is that we should choose the distribution of resources that satisfies:

$$\arg\sup_{\varphi} \int_0^1 min_i v^i\left(\pi, \varphi^i\right) d\pi$$

That is to say: We begin by looking at the minimum level of well-being achieved by those in each effort-percentile across all types. We then choose the distributive scheme that maximizes the average of those minimum values from each

[19] The view is developed in (Roemer 1994, pp. 276–301).

effort-percentile across all types. The justification for this is as follows. We cannot reasonably expect to see the same absolute amount of effort exerted by an average member of a disadvantaged type as is exerted by an average member of an advantaged type. It would not be reasonable for the member of the disadvantaged type to exert the same absolute level of effort as the member of the advantaged type, since society effectively tells him that his efforts will (in all likelihood) not be rewarded past a given point, which is lower than the point the member of the advantaged type can reasonably hope to reach. Roemer directs our attention, then, to the relative degree of effort exerted by an individual compared with the effort exerted by other individuals within the same type. And his central message is that we should redistribute resources in such a way that any two individuals who exert an average degree of effort *for their own types* should receive the resources they require in order to attain equal levels of well-being. Or, if there is a distribution which will maximize the average well-being of the worst-off members of each effort-percentile across types, but which differs from the distribution that equalizes the well-being of members of each effort-percentile across types, we should prefer the maximining distribution. In this way, we distribute resources so that the level of well-being an individual is able to achieve for himself depends only on his degree of autonomous effort, and he has the opportunity to achieve the highest feasible level of well-being given his degree of autonomous effort. A more accurate label for Roemer's approach would thus be "maximining well-being subject to degree of autonomous effort."

Roemer's view differs in an important way from the equality of well-being subject to desert view discussed above. The crucial difference is his focus on the degree of autonomous effort made by an individual within a type, and the suggestion that individuals who belong to the same effort-percentile across types should receive the resources they need in order to have an equal chance at achieving an equal level of well-being, despite the fact that their absolute levels of effort may differ. What was objectionable about the sort of view considered above was that it tied desert to an individual's absolute level of effort, and was not sensitive to the fact that morally arbitrary inequalities affect the level of effort which it would be reasonable for an individual to make. Roemer's approach avoids this problem by compensating for those undeserved disadvantages, and tying that compensation to the degree of effort exerted by an agent as compared with other members of his own type.

As it stands, Roemer's model suffers from an important and obvious defect. It does not take into account accepted risk. Within a given type, two agents may work equally hard, but one may choose to enter into riskier ventures than another. If the risk is accepted, or avoided, freely, knowingly, and with access to the available relevant information, and the outcomes of the two choices differ, this difference is not one that a theory of distributive justice should be concerned with. The fact that Roemer's model is insensitive to accepted risk is, I suspect, the problem at the heart of a number of criticisms his view has received.[20] It suffices for my purposes to

[20] See, for example, the responses to Roemer's contribution to the Boston Review's "Social Equality and Personal Responsibility" forum, particularly the responses of Arthur Ripstein, Susan Hurley, and T.M. Scanlon (Roemer et al. *Boston Review,* Vol. 20 No. 2, 1995).

register this concern as the source of a needed amendment to Roemer's proposal. We must look not only at the levels of well-being achieved by the members of each effort-percentile across types, but also at the risks they freely, and knowingly accept—careful all the while to keep in mind that those in disadvantaged types will be more likely to take on risks within an atmosphere of social pressure, without knowing exactly what risks they are taking on, or without having access to all the available information relevant to that decision. We should then turn our attention to the minimum level of well-being achieved within a particular effort-percentile and a particular risk-acceptance-percentile across all types.

In the remainder of this section, I consider two objections to Roemer's view. The first, made by Mathias Risse, charges that Roemer's theory both commits to and is inconsistent with a compatibilist position on determinism and responsibility. I will argue that Risse's objection rests on a mistake.[21] The second is my own objection to the adequacy of Roemer's theory as an aspirational account of distributive justice. I will argue that it is instead a transitional account, and one which, as we will later see, complements my own aspirational account nicely.

Roemer explicitly commits himself to a compatibilist position with respect to determinism and responsibility. More specifically, he adopts a Scanlonian view of responsibility, and cites Scanlon's own account of what is and is not required for an individual to be responsible for his own decisions and actions:

> What is required [for responsibility] is that what we do be importantly dependent on our process of critical reflection, that the process itself be sensitive to reasons, and that later stages of the process be importantly dependent on conclusions reached at earlier stages. But there is no reason, as far as I can see, to require that this process itself not be a causal product of antecedent events and conditions. (Scanlon 1988, cited in Roemer 1998, p. 17)

The Scanlonian account of individual responsibility is the one that I accept as well. This is why I tie responsibility to the possession of the capacity for autonomy, and the freedom to exercise that capacity and act on one's autonomous conclusions. My account of autonomy just is an account of the process of critical reflection which Scanlon identifies as the basis for responsibility—an account which is far more rigorous and precise than any other currently on offer.

Let us look, then, at Risse's argument for the claim that one who holds Roemer's view of distributive justice cannot also be a compatibilist. For if Risse turns out to be right, I will not be able to adopt Roemer's view as my preferred transitional theory while maintaining a Scanlonian view of responsibility. According to Risse, the problem for Roemer's view arises in virtue of his assumption that once we have identified all the external circumstances that affect an individual's effort for which he is not responsible, we will see identical distributions of degree of effort across types. Risse claims that only a believer in libertarian free will should expect to see this. I quote his argument in full:

[21] A number of other philosophers have made a variety of objections to Roemer's view, but he has done an excellent job of answering them (Roemer 2003).

> Libertarians hold that choice is uncaused and, thus, that there is no causal relationship between those aspects of a person's condition for which she is responsible and those for which she is not.
>
> Therefore, they also deny that there is any correlation between effort distribution curves and types. For if there were such a correlation, it would presumably have to be explained by a common cause, which would conflict with the libertarian idea of uncaused choice. So according to the libertarian idea of uncaused choice, effort curves will be independent of types, and thus for large types, those curves will be identical "almost certainly," as probability theorists say. (Risse 2002, p. 729)

A compatibilist, on the other hand, is supposed to have no grounds for expecting identical distributions across types, a phenomenon that Risse refers to as "No-Variance":

> Yet compatibilists reason differently. They acknowledge that aspects of a person's condition for which she is responsible are themselves caused. If choices are caused, they will ultimately be caused by or at least be correlated with aspects of a person's condition for which she is not responsible (her circumstances). Thus for a compatibilist to accept No-Variance would be to accept that, although she conceives of choices as caused, the effort distribution curves are shaped as if the set of possible causes (her circumstances) were irrelevant to it, that is, as if choices were uncaused. Thus not only is a compatibilist (unlike a libertarian) lacking any positive reason for finding No-Variance more plausible than any other claim about the shape of the distribution curves; but, what is more, in virtue of rejecting the idea of uncaused choice, a compatibilist also finds it immensely plausible that the effort distribution curves will vary across types. That is just what it is to be a compatibilist. (A compatibilist would also be keenly aware that the acceptance of any other thesis about the shape of the distribution curves would lead Roemer's theory to entirely different policies; it may well be true that most any policy could be the recommendation of Roemer's theory given a suitable thesis of that sort.) So a compatibilist has merely an incredulous stare for a theorist who asks her to endorse No-Variance over other theses about the shape of the effort curves. (Risse 2002, p. 729)

Risse's mistake is his claim that the compatibilist can only accept No-Variance if he accepts that effort distribution curves are shaped as if choices are uncaused. To see why this is wrong, let us begin with a simple example. Suppose we have a fair roulette wheel. We also have large containers of metal roulette balls. The balls in each container are identical. The balls in one container differ from those in another only in being a different color, which difference has no impact on the way they interact with the wheel. Let us call balls of one color "type 1" balls, balls of another color "type 2," and so on. We play each ball of each type, and record the slot it falls in, so that we have a distribution associated with each type. The type and the distribution are independent—being of type 1 makes it neither more nor less likely that a ball will fall in a given slot than being of type 2 would. Given enough balls of each type, we will in the long run see identical distributions. Playing the balls of the various types thus yields a set of independent identical distributions.

Now here is the important point: *The roulette balls do not possess libertarian free will*. For any given play, a fully deterministic process determines which slot the ball falls in. Roemer's point is that there are aspects of our circumstances that both determine our absolute levels of effort *and* affect our responsibility for those levels of effort, and there are aspects of our circumstances that determine our levels of

effort *without* affecting our responsibility for those levels of effort. When we have taken all the former into account, so that we are rightly (on the compatibilist view) held responsible for where we fall on the typed effort distribution, we are still left with the latter, and these latter aspects of our circumstances do causally determine our levels of effort. But at this point, the effort distribution for each type is independent of the type to which it belongs. Given an individual plucked from a particular type, what type he belongs to has nothing to do with our expectation regarding where on the effort distribution curve he will fall. This is in fact the criterion of success for defining the types. If almost all of the members of an advantaged type are found together at the upper edge of the range of effort found for members of that type, and almost all the members of a disadvantaged type are found together at the lower edge for their type, then the types have not taken into account all the external features which relevantly affect agents' output of effort. We know we have defined types properly when such discrepancies disappear, and the members of one type are no more likely to be found in a given effort-percentile than the members of another—the distributions will be type-independent. And we should expect the effort distributions to be identical across types. If they were not identical, that would be a sign that some factor was differentially affecting effort output within different types, and that factor ought to be accounted for in the definitions of the types.

The source of Risse's mistake is not difficult to identify. Within each type, effort is a random variable. It is a popular view among some philosophers that there are no deterministic probabilities. That is, it is never the case that we must, in the final analysis, ascribe probabilities to the types of outcomes that result from deterministic processes. But this is a mere metaphysical prejudice; and as usual in such cases, the cure is greater familiarity with science. Classical statistical mechanics (CSTM) is an extremely successful and important deterministic theory that does make the assumption that probabilities must be ascribed to the types of outcomes of the processes it models. As Barry Loewer explains: "The probability assumption of CSTM is essential to its predictions, explanations and laws…The micro-canonical [probability] distribution is required to account for all thermodynamic phenomena: the diffusion of gases, the melting of ice in water, and a myriad of other time-asymmetric phenomena" (Loewer 2001, p. 610). Deterministic probabilities are also used in quantum mechanics with respect to small systems, and in population genetics (Strevens 2011). Risse is free to make the move, popular among philosophers, of claiming that the use of probabilities by deterministic theories is a sign of their imperfection and incompleteness, and that any final deterministic theory will exclude probabilities. But again, this seems little more than an expression of metaphysical prejudice. What makes one a naturalist is the fact that one takes the best science of the time as one's guide to understanding the world. To insist that science will, eventually, given sufficient resources and time, vindicate one's own metaphysical preferences, is to play the role of fortune-teller rather than philosopher. There is no sound basis, then, for Risse's assertion that type-independent effort distributions portray effort as uncaused.

Nor is there any basis whatsoever for making any claim about what effort distribution curves would look like if the agents exerting the effort possessed libertarian

free will. The idea that these will be identical if the types are sufficiently large results from another widespread philosophical mistake. Risse believes—mistakenly, as we have seen—that Roemer's effort distribution curves are a sign that effort is the result of a stochastic process, as opposed to a deterministic one. This is a sufficient condition (though again, as we have seen, *not* a necessary one), for identical distributions given large enough types. But even if they were the result of a stochastic process, this would not show that effort is uncaused, in the sense of being the product of acts of libertarian free will. It would show, instead, that effort resulted from processes that were, at least in part, random. But—and here is the mistake—stochastic (indeterministic, random) processes have nothing whatsoever to do with libertarian free will.[22]

In fact, Roemer need not commit himself to strict determinism at all; he could endorse the view that the universe is fundamentally stochastic, that some outcomes result not from deterministically sufficient cause, but from a combination of deterministic cause and objective chance.[23] Because stochastic processes and the exercise of libertarian free will have nothing to do with one another, Roemer is free to adopt a stochastic version of the Scanlonian account of personal responsibility. And he can of course make all the same claims about identical independent effort distributions while identifying the values that constitute those distributions as outcomes of a stochastic process, rather than a deterministic one that makes use of probabilities. Assigning probabilities to the values taken on by a random variable, moreover, only makes sense given the assumption that the value the variable takes on in a particular instance results from either a stochastic process or a deterministic one that makes use of probabilities. The fact is that there is no basis for saying anything whatsoever about what effort distribution curves would look like if effort were the result of act of libertarian free will. What Risse believes follows from compatibilism actually follows from libertarianism. Risse's critique of Roemer is exactly wrong.

Roemer actually assumes that we would see not an equal distribution of degrees of effort—as we would see an equal distribution on a fair roulette wheel—but rather a Gaussian distribution. That is, within any given type, many people will fall in the 50th percentile, with fewer and fewer as we move out toward the 1st and the 99th. This is a common assumption for the distribution of any sort of measurable human behavior. A closer analogy would thus be a "Gaussian roulette wheel," with a large slot at 50 and progressively smaller slots as we move out to 1 and 99. Given enough individuals within each type—enough plays on the Gaussian wheel—we should expect to see a set of identical independent Gaussian distributions, just as Roemer assumes we will. And at no point do we have to assume that humans—or roulette balls—have libertarian free will. All we need do is make the (reasonable) assumption that relative degree of autonomous effort is a random variable distributed normally throughout the total population. To reject No-Variance—to hold that some

[22] For a particularly illuminating discussion of the incompatibility of libertarian free will with both deterministic and indeterministic physical laws, see (Nayakar and Srikanth 2010).

[23] This is the view famously expounded by Peirce, under the name *tychism*.

groups simply and ineliminably have a higher proportion of members who work hard by the standards appropriate to the circumstances of the group—amounts to prejudice. And the compatibilist can reject prejudice just as well as the libertarian.

So much for Risse's objection. Let us conclude, then, by considering whether Roemer's theory is adequate as an aspirational account of justice. It seems to me that it is not, for one simple reason. It makes individuals dependent on the State *in the wrong way* for the provision of the opportunity for leading a good life to which all are equally entitled. Consider an example. Suppose we have a society that currently has well-crafted anti-discrimination laws that are effectively enforced. There might nonetheless be very good reason within such a society to incentivize affirmative action hiring and school admission practices. This might be so if a particular ethnic group has historically been marginalized and disadvantaged within the society. But that hardly shows that affirmative action policies would be part of a truly just society. In fact, it is hardly plausible that they would be. The reason is that in a truly just society, we should expect affirmative action practices to be unnecessary. This is not a utopian aspiration. It does not require that prejudice itself be eliminated. It does not even require that enforcement of anti-discrimination laws be perfect. What it does require is the creation of the social conditions in which members of the historically marginalized group have the same opportunities for self-development as those in historically advantaged groups and the same incentives to exert absolute levels of effort comparable to those associated with the members of historically advantaged groups. These conditions can be realized despite the persistence of a limited amount of prejudice within a society, and the occasional unpunished violation of anti-discrimination laws. Affirmative action policies are valuable insofar as they are part of a transitional scheme. Their goal is to bring us closer to becoming a society in which the new generations of groups that were historically marginalized have those equal opportunities for self-development and incentives for effort.

Roemer's model of equality of opportunity for welfare is like affirmative action writ large. It is a distributive scheme that seeks to equalize (or maximin) the level of well-being that individuals who exert the same relative degree of effort are able to achieve. This lifting-up of those who find themselves, for reasons beyond their control, in disadvantaged groups is precisely what is required to move a fundamentally unjust society closer to being just. It is what is required to create, gradually over successive generations, the conditions in which all individuals enjoy the equal opportunities for self-development and incentives for effort which they deserve. And we will see that there is reason to think that it is a transitional account which dovetails well with my own aspirational one. But it is just that: a transitional theory. Like affirmative action, there is no place for Roemer's distributive policies within a truly just society.[24] We would expect a just society to focus on maintaining and

[24] The same can be said, I think, of the theory of distributive justice developed by the economist and philosopher Marc Fleurbaey (2009). Fleurbaey's goal is similar to Roemer's: he aims to articulate and defend a detailed and rigorous egalitarian theory which is sensitive to the moral importance of autonomous effort and personal responsibility. In its details, his theory differs from Roemer's in a

preserving conditions of equal opportunity and incentive for effort for all, rather than on equalizing chances for well-being by compensating those who have been denied those very conditions.

References

Arneson, R. 1998. Real freedom and distributive justice. In *Freedom in economics: New perspectives in normative analysis*, ed. J.-F. Laslier et al., 165–197. New York: Routledge.
Arneson, R. 1999. Against Rawlsian equality of opportunity. *Philosophical Studies* 93(1): 77–112.
Arneson, R. 2000. Perfectionism and politics. *Ethics* 111(1): 37–63.
Arneson, R. 2009. Value pluralism does not support liberalism. *San Diego Law Review* 46(4): 925–940.
Barry, B. 1988. Equal opportunity and moral arbitrariness. In *Equal opportunity*, ed. N. Bowie. Boulder: Westview.
Charles, D. 1988. Perfectionism in Aristotle's political theory. In *Oxford studies in ancient philosophy, supplementary volume*, 185–206. Oxford: Oxford University Press.
Cohen, G.A. 1993. Equality of what? In *The quality of life*, ed. M. Nussbaum and A. Sen, 9–29. Oxford: Clarendon Press.
Cohen, G.A. 1997. Where the action is: On the site of distributive justice. *Philosophy and Public Affairs* 26(1): 3–30.
Cohen, G.A. 2008. *Rescuing justice and equality*. Cambridge, MA: Harvard University Press.
Dworkin, R. 1981. Equality of what? Part II: Resources. *Philosophy and Public Affairs* 10(4): 283–345.
Dworkin, R. 1985. *A matter of principle*. Cambridge, MA: Harvard University Press.
Dworkin, R. 1987. What is equality? Part 3: The place of liberty. *Iowa Law Review* 73(1): 1–54.
Dworkin, R. 2000. *Sovereign virtue: The theory and practice of equality*. Cambridge, MA: Harvard University Press.
Fleurbaey, M. 2009. *Fairness, responsibility, and welfare*. Oxford: Oxford University Press.
Frankfurt, H. 1988. Alternative possibilities and moral responsibility. In *The importance of what we care about*, 1–10. Cambridge: Cambridge University Press.
Galston, W.A. 1991. *Liberal Purposes*. Cambridge: Cambridge University Press.
Galston, W.A. 2009. The idea of political pluralism. In *Moral universalism and pluralism*, NOMOS, vol. XLIX, ed. H.S. Richardson and M.S. Williams. New York: NYU Press.
Green, T.H. 1883. *Prolegomena to ethics*. Oxford: Clarendon Press.

number of ways; it would hardly profit us to undertake the task of working out the relationship between the two theories here. Fleurbaey's theory does have much to recommend it as a transitional theory of justice, but I have chosen to focus on Roemer's theory because, at the end of the day, I think it superior. Fleurbaey's theory has a significant vice which does not make it particularly suited to play the role of transitional theory with respect to *any* aspirational theory that requires—as mine does, and as I think any plausible one must do—profound institutional change compared with the *status quo* in contemporary capitalist democracies. It assumes the existence of a "natural" allocation of resources—the *laissez-faire* allocation—and allows deviations from this allocation only for the sake of compensating individuals for losses/disadvantages due to circumstances beyond their control. But as Roemer points out, whatever it is Fleurbaey means by the *laissez-faire* allocation "there is, of course, *some* set of social institutions which determines those outcomes: and why should that set deliver justice, after the effect of circumstances has been neutralized?… [Fleurbaey's theory] assigns too much moral legitimacy to extant institutions" (Roemer 2011, p. 133). This point—the fact that there is no "natural" allocation outside any set of social institutions—will be developed in the next chapter.

Kenny, A., and C. Kenny. 2006. *Life, liberty and the pursuit of utility*. Charlottesville, VA: Imprint Academic.
Lerner, A. 1944. *The economics of control*. New York: Macmillan.
Lippert-Rasmussen, K. 1999. Arneson on equality of opportunity for welfare. *Journal of Political Philosophy* 7(4): 478–487.
Loewer, B. 2001. Determinism and chance. *Studies in the History of Modern Physics* 32(4): 609–620.
Machin, D. 2013. Rawls's difference principle as compensation for social immobility. *Political Quarterly* 84: 4.
McCarthy, D. 2006. Utilitarianism and prioritarianism I. *Economics and Philosophy* 22(3): 335–363.
McCarthy, D. 2008. Utilitarianism and prioritarianism II. *Economics and Philosophy* 24(1): 1–33.
Nayakar, C.S.M. and R. Srikanth. 2010. *Quantum randomness and free will*. Available at http://arxiv.org/pdf/1011.4898v1.pdf.
Nussbaum, M. 1988a. Aristotle on political distribution. In *Oxford studies in ancient philosophy, supplementary volume*, 145–184. Oxford: Oxford University Press.
Nussbaum, M. 1988b. Reply to David Charles. In *Oxford studies in ancient philosophy, supplementary volume*, 207–214. Oxford: Oxford University Press.
Nussbaum, M. 1990. Aristotelian social democracy. In *Liberalism and the good*, ed. R.B. Douglass, G. Mara, and H. Richardson, 203–252. New York: Routledge.
Nussbaum, M. 1993. Non-relative virtues: An Aristotelian approach. In *The quality of life*, ed. M. Nussbaum and A. Sen, 242–269. Oxford: Clarendon Press.
Parfit, D. 1997. Equality and priority. *Ratio* 10(3): 202–221.
Plato *The Republic*. Trans. A. Bloom. 1991. New York: Basic Books.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Rawls, J. 1987. *Political liberalism*. Cambridge, MA: Harvard University Press.
Rawls, J. 1999. *The law of peoples*. Cambridge, MA: Harvard University Press.
Rawls, J. 2001. *Justice as fairness*. Cambridge, MA: Harvard University Press.
Risse, M. 2002. What equality of opportunity could not be. *Ethics* 112(3): 720–747.
Roemer, J. 1982. *A general theory of exploitation and class*. Cambridge, MA: Harvard University Press.
Roemer, J. 1994. *Theories of distributive justice*. Cambridge, MA: Harvard University Press.
Roemer, J. 1998. *Equality of opportunity*. Cambridge, MA: Harvard University Press.
Roemer, J. 2003. Defending equality of opportunity. *The Monist* 86(2): 261–282.
Roemer, J. 2011. Review: Fairness, responsibility, and welfare. *Journal of Economic Inequality* 9: 129–135.
Roemer, J., et al. 1995. Social equality and personal responsibility. *Boston Review* 20(2).
Scanlon, T.M. 1988. The significance of choice. In *The Tanner lectures on human values*, vol. 8, ed. S. McMurrin. Salt Lake City: University of Utah Press.
Scitovsky, T. 1976. *The joyless economy*. Oxford: Oxford University Press.
Smith, A. 1776/2003. *The wealth of nations*. New York: Bantam.
Strevens, M. 2011. Probability out of determinism. In *Probabilities in physics*, ed. C. Beisbart and S. Hartmann. Oxford: Oxford University Press.
Wall, S. 1998. *Liberalism, perfectionism and restraint*. Cambridge: Cambridge University Press.

Chapter 10
Beyond the Old Economics

1 Introduction

This chapter introduces the economic theory which serves as the background for the theory of distributive justice developed and defended in the next chapter. The realization of social justice as I conceive of it—a society in which every individual enjoys an equal share of liberty, as I have defined it—will require the existence of a strong State whose institutions both shape and participate in the market. Any attempt to establish and preserve equality of liberty (in my sense) will be undermined by the emergence of the large disparities in wealth and economic power which market economies can, and often do, give rise to. By "economic power," I mean the ability of individuals (in virtue of their wealth), or firms (in virtue of their wealth, market share, or systemic importance to a national economy), to shape the operations of the marketplace and public policy to suit their own interests, either directly or through exercising influence over political, legislative, judicial and regulatory processes. As we will see, these large disparities threaten the freedom of many to develop and exercise their autonomy, while vastly expanding the freedom and power of few. Without public institutions and policies that prevent these large disparities, the goal of equal liberty is unrealistic. But any egalitarian theory of social justice is sure to encounter objections from those, steeped in orthodox economic theory, who claim that State "interference" (for so they would term it) in the market is bound to bring inefficiency, instability, and stagnation.

The primary purpose of this chapter is therefore to show that this sort of objection is groundless. I begin with a discussion of my reasons for rejecting General Equilibrium Theory, the heart of economic orthodoxy, as an adequate framework for understanding the workings of any actual economic system. I then discuss my commitment to the Evolutionary-Institutional school of economic theory, and describe what I take to be its core claims and main insights. I highlight the various points at which taking this perspective on economic systems lends support to broadly egalitarian policies. Achieving the economic goals of stability, efficiency, and growth will turn out to require strong public institutions which actively limit the

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_10

emergence of large disparities in wealth and economic power—just the sort of public institutional action which is a prerequisite for equality of liberty. This discussion of economic theory also provides us with all the tools we need to quickly dispense with one family of arguments for a free market and a minimal State—that is, a market which is largely free from the influence of public institutions, the role of these being limited to the establishment and enforcement of contractual and property rights.[1] This first, axiological family of arguments appeals to the supposed benefits of this arrangement in the forms of economic stability, efficiency, growth, and innovation. I follow this with a brief review of the deontic arguments for the free market and the minimal State, which are based on considerations of individual desert and self-ownership. We will see that none of these arguments stand up to scrutiny.

2 The Limits of General Equilibrium Theory

2.1 New Keynesian Economics and the Role of Policy

The basic claim of General Equilibrium Theory (GET) is that, if there is a perfectly competitive market (in a rather idiosyncratic sense of "competitive" which we will explore below) for every commodity, then each market will arrive at a price for its commodity at which excess demand (demand for the commodity above what those who supply it are willing to bring to the marketplace) vanishes to 0. Following the work of Kenneth Arrow and Gerard Debreu in the mid-1950s (Arrow and Debreu 1954), GET "became the fundamental framework for theoretical discourse" in economics (Ackerman 2002, p. 119). How central it remains to the discipline is a matter of some debate, as most present-day professional economists will admit that the theory has many serious flaws—some of which will be discussed below—although much current professional economic research still takes place within the general equilibrium framework. What should be less controversial is the claim that the results of GET still exert a dominant influence on economic policy analysis and recommendations. Economists "often talk as if deductions from general equilibrium theory are applicable to reality" (Ackerman 2002, p. 121). The theory "is widely cited in a normative context, often in textbooks or semitechnical discussion, as providing the rigorous theoretical version of Adam Smith's invisible hand and demonstrating the desirable properties of a competitive economy" (Ackerman 2002, pp. 119–120). The first fundamental theorem of welfare economics states that every general equilibrium results in a Pareto-efficient allocation of goods. This is the point which is generally taken to be a rigorous confirmation of Adam Smith's claim that competitive markets allocate goods efficiently. The macroeconomic models of both

[1] A free market and a minimal State—where the former is simply the type of market created by the institutions of the latter, as all markets are created by a set of public institutions of one type or another—should be understood as two elements of a single entity: what Karl Polanyi called "the Market Society" (Polanyi 1944).

New Keynesian and New Classical economists—the theoretical foundations of both liberal and conservative approaches to economic policy, respectively—are general equilibrium models.

The crucial assumption required by GET and the first fundamental welfare theorem to show that the market reaches an efficient equilibrium is that market competition is perfect. This is an assumption with many parts.[2] The one that has received the most attention is the assumption of perfect information. In a perfectly competitive market, every participant is assumed to know the exact quality and the equilibrium price of the commodity offered in that market, and to have incurred no costs in acquiring this knowledge. The real world is obviously not a world of perfect information, and we can expect some degree of market inefficiency as a result. Given inefficient markets which suffer from imperfect information (perhaps among other problems), do we have reason to believe that there are often forms of State intervention available that will be Pareto-improving, owing to the fact that they will produce a net increase in efficiency by improving the distribution of information? It turns out the answer to this question is in the affirmative. The result that established this is now called the Greenwald-Stiglitz theorem, after the New Keynesian economists Bruce Greenwald and Joseph Stiglitz, its authors (Greenwald and Stiglitz 1986). As Stiglitz explains:

> [W]henever information is imperfect or markets (including risk markets) are incomplete—that is, essentially almost always—competitive markets are not constrained Pareto efficient. Taking into account the cost of improving information or creating markets, some individuals could, in principle, be made better off without making anyone else worse off. (Stiglitz 2000, p. 1458)

And these improvements can often be attained through available courses of interventionist State action:

> [I]n many cases, not only can it be demonstrated that there exist Pareto-improving government interventions, but also that the kind of intervention required can be simply related to certain parameters that, in principle, are observable. (Greenwald and Stiglitz 1986, p. 230)

There will be room for Pareto-improving State intervention so long as there are negative effects of market inefficiencies whose removal more than outweighs the deadweight loss created by the market distortion introduced by a tax or regulation (Greenwald and Stiglitz 1986, pp. 237–238). We will see in the next chapter that creating a better distribution of information—so that individuals have greater knowledge of what risks they may be exposed to, and better access to the available information relevant to their risk-taking decisions—is crucial for distributive justice. The new economics of information gives us good reason to believe that doing this is possible.

[2] Specifically: (1) infinite buyers and sellers; (2) no market entry/exit barriers; (3) instant and costless mobility of the factors of production; (4) no increasing returns to scale; (5) all commodities in a given market are exact substitutes; (6) all firms are profit-maximizers; (7) all consumers are expected utility-maximizers; (8) complete consumer knowledge of prices, utility, quality and production methods of all commodities; (9) costless exchange; (10) no externalities from exchange.

Another imperfection in market competition, at least according to the New Keynesian school, is the existence of price- and wage-stickiness. To call prices and wages "sticky" is to claim that when a market experiences a shock which shifts the equilibrium price (or equilibrium wage, as the wage is the price of labor) away from the current price, firms do not immediately and costlessly adjust to the new market conditions by moving to the new equilibrium price. This fact leads to market inefficiencies; and again, State action—in the form of monetary or fiscal policy action—can lead to a net gain in efficiency. Stickiness is perhaps the central preoccupation of New Keynesian theory. One way in which information can be imperfect is by being sticky: participants in a market do not instantly and costlessly update their knowledge with changing market conditions.

New Keynesian theory thus offers at least one line of argument against the claim that the goal of economic efficiency is best served by a minimal State. Real markets are not perfectly competitive; prices, wages and information are all sticky.[3] Governments can exercise monetary and fiscal policy, change the distribution of information, and create markets that fail to emerge on their own. And by doing so, they can create net increases in market efficiency. This line of argument does not challenge GET itself. It is still assumed that if competition were perfect in all markets, an equilibrium would be reached, and this equilibrium would be Pareto-optimal. It is also assumed that if an economy were to experience a shock only once in a great while, it would eventually return to equilibrium without any help in the form of State action. Stickiness slows things down, but does not prevent return to equilibrium all on its own. But the problems with GET run much deeper.

2.2 The Question of Stability: The Sonnenschein-Mantel-Debreu Theorem

The Sonnenschein-Mantel-Debreu theorem has been the central result on the limitations of GET since its discovery and refinement in the mid-1970s (Sonnenschein 1973; Mantel 1974; Debreu 1974). The theorem concerns the precise nature of the relationship between the properties of the excess-demand functions of individual consumers and of the economy as a whole (the aggregate of all consumers). It establishes that individual consumer demand functions can only be aggregated into a market demand function with the same tidy mathematical properties as the individual functions under certain specific, restrictive, and unrealistic assumptions. The technical details need not detain us here—though we will consider some of the required assumptions below. What really matters are two implications of the theorem. First, the theorem shows that, given standard assumptions about the preferences, behavior, and constraints of consumers and suppliers, there is no unique answer to the question of what equilibrium price-level and supply-/demand-level an

[3] New Keynesian models also allow for the existence of a handful of other market imperfections, such as the existence of monopolies.

economy will converge on. The second implication is that "almost any continuous pattern of price movements can occur in a general equilibrium model… Cycles of any length, chaos, or anything else you can describe, will arise in a general equilibrium model for some set of consumer preferences and initial endowments" (Ackerman 2002, p. 122). It is possible, in other words, that an economy will *never* find its way to an equilibrium price-level, even in the absence of exogenous shocks (like sudden crop failures) or externally imposed constraints (such as price-ceilings). It is also possible that, if an economy does find its way to an equilibrium, that equilibrium will be unstable, in the sense that it will not be able to return to an equilibrium after a small shock.

There are ways of mitigating this second possibility, by making assumptions which will guarantee a return to equilibrium after a small shock, assuming the pre-shock state was an equilibrium. In a pure exchange economy (one in which no commodities are produced), the required assumption is gross substitutability: that for every consumer, every commodity is a substitute for every other commodity. This implies that when the price of one commodity rises, no one continues to buy the same quantity of that commodity. Whether or not one finds this assumption acceptable, it is not sufficient to resist the implications of S-M-D in a production economy (Kehoe 1980). In the context of a production economy, the problem is solved in one of two ways. One is by assuming that the economy can be modeled as if there were only one consumer and one producer—the so-called representative individual or household, and the representative firm—whose behavior coincides with the aggregate behavior of heterogeneous consumers and firms. This is the approach of both New Keynesian and New Classical macroeconomics. But strict and unrealistic assumptions about individual consumer preferences are required to generate an aggregate excess demand function that behaves as if it belongs to a single utility-maximizing agent. In particular, we must assume that individuals have *homothetic* preferences, which implies that given the price-level, individuals with different incomes will demand goods in exactly the same proportions (though obviously in different quantities).[4] The other is by using a one-commodity model of production. This may be taken literally, in the sense of assuming that there is only one type of thing that is consumed, and this same type of thing is the only type of thing used in the process of production. Or it may equivalently be interpreted as assuming that all capital and consumption goods are made of the same stuff, "putty," which is homogenous and perfectly, instantly, costlessly malleable (Cohen and Harcourt, pp. xli–xliii). This is essentially a substitutability assumption parallel to the gross substitutability assumption of the pure exchange model, but on the productive side: an assumption that producers can instantly, costlessly switch to a new type of capital good when the price of the type they use rises. Both of these assumptions are

[4] For a number of trenchant criticisms of representative agent theory, see (Kirman 1992). The homothety assumption, as we will see, amounts to a rejection of the very cornerstone of Keynes' theory—the psychological law of the diminishing marginal propensity to consume, which Keynes calls the key to the whole problem of stable but sub-optimal levels of employment (Keynes 1936/1964, pp. 19–20).

obviously wildly unrealistic. Their use is justified by the claim that despite being so far removed from reality, the models that employ them will nonetheless be capable of making accurate predictions of economic behavior, and the question of why they are able to do so is not one that should worry us.[5] Even though consumers are not identical and capital goods are not made of putty, it is good enough to model the economy as if they were. Anyone who is skeptical of this claim will find himself with an embarrassment of riches. It is a matter of simple historical fact that the predictive track record of modern economic theory comes nowhere near to supporting this confidence.[6]

But there is an even more fundamental problem. The first possibility—that starting from initial conditions (i.e. initial preferences, technological limits, and pretrade endowments), an economy of utility-maximizing agents free from externally imposed constraints will not ever converge to an equilibrium in the first place—remains. The impact of the S-M-D theorem, then, is that it leaves the claim that production economies, at least in the absence of large exogenous shocks and externally imposed constraints, eventually find their way to an initial equilibrium, based on nothing but faith, "the faith…that disequilibrium dynamics will converge to equilibrium outcomes…" (Cohen and Harcourt 2005, p. xlviii). But as we will now see, this article of faith has become untenable.

2.3 The Dynamics of GET

We have identified an article of faith: in the absence of exogenous shocks and externally imposed constraints, an economy beginning from initial conditions will converge on an equilibrium, and will return to an equilibrium after a (not too large) shock does occur. Some economists believe that the existence of price, wage and information stickiness can cause the latter process to take longer than it otherwise would. Some of them believe the government can take effective action to speed the return to equilibrium in such a scenario. But the central belief shared by both camps is that forces exist in disequilibrium which will eventually push an economy to equilibrium. It turns out, however, that we can go a step beyond the S-M-D theorem, and show that disequilibrium dynamics which result in anything but a stable equilibrium are not only possible for and consistent with GET; they are virtually all we can expect from it.

At this point we need to distinguish two types of process by which an economy might converge on an initial equilibrium. The first is the Walrasian *tâtonnement* process, in which we imagine an auctioneer calling out possible prices for each commodity so that suppliers and consumers can show how much of the commodity

[5] This is Milton Friedman's famous argument that it is appropriate for economics to use an 'as-if' methodology (Friedman 1953).

[6] For a recent study of the inability of the most sophisticated macroeconomic models to forecast the business cycle, see (Edge and Gurkaynak 2011).

they would be willing to supply/demand at that price, but not setting the price of any one until excess demand equals zero. No economic activity takes place until the equilibrium price for each commodity is found. Non- *tâtonnement* processes allow economic activity to start immediately, and the price-level adjusts over time based on market forces. The S-M-D theorem shows that the *tâtonnement* process in a pure exchange economy can fail to end in a stable equilibrium. This result is extended by the work of Donald Saari, who showed how demanding our assumptions about a non- *tâtonnement* price-adjustment mechanism for a pure exchange economy must be in order to be guaranteed to converge on an equilibrium (Saari 1985). What I will discuss now is the case closest to reality, that of a production economy with a non-*tâtonnement* price-adjustment mechanism. And we will see that the source of the trouble is one of the most fundamental assumptions of the general equilibrium enterprise.

In order to understand why this article of faith is untenable in the most realistic case, we have to go back to the origins, both logically and historically, of GET. The purpose of GET is to demonstrate that a stable equilibrium is reached when every economic participant is maximizing his own utility, subject to his budget/production constraint, in the absence of exogenous shocks and externally imposed constraints. Demonstrating this requires that we be able to define utility functions for economic agents. If one thinks of preferences as psychological objects, it is a simple matter to define utility functions by imposing the necessary restrictions on preferences over bundles of commodities—the sorts of restrictions discussed in Chap. 3. But GET must assume that agents *always act* so as to maximize their utility (agents in GET models do not fall victim to weakness of will), and thus utility functions must be derivable from observable behavior. In fact, the belief that utility functions are derivable from observable behavior is what convinced the early neoclassical economists that economics could be a genuine science, a discipline wholly separate from philosophy and psychology.

The road to undermining the faith that disequilibrium dynamics eventually brings the economy to equilibrium begins with a crucial assumption required to derive utility functions from observable behavior. In order to explain the assumption, I will draw on Philip Mirowski's translation of and commentary on a famous letter written in 1900 to Léon Walras, one of the founders of neoclassical economics, by the physicist and mathematician Hermann Laurent (Mirowski 1989, pp. 245–246). Both the interpretation of Laurent's argument, and the validity of the much grander claims which Mirowski uses it to support, are controversial. My own interpretation of Laurent's argument differs slightly even from Mirowski's, although I imagine he should be even more sympathetic to it than to his own. I will briefly address the controversies below. But first, let us get clear on the assumption in question.

Suppose we have a consumer, who has consumed quantities of merchandise dq_1, dq_2,…dq_n during time period dt. The total expenditure dE made by this consumer during dt is then

$$dE = p_1 dq_1 + p_2 dq_2 + \ldots + p_n dq_n \tag{10.1}$$

where p_i is the price of one unit of commodity i. Laurent does not actually name the scalar function here, but simply refers to the expression on the right as the 'total price.' From the context, it is clear that by this he means the total expenditure made during dt. This is a bit of observable consumer behavior: this person bought these commodities in these quantities at these prices during this time. In order to derive a utility function for this consumer, Laurent points out, we need to make an assumption about (10.1). We need to assume that it is an *exact differential equation*. This assumption, as Mirowski notes, is equivalent to the assumption that the price vector $\mathbf{p} = (p_1, p_2, \ldots)$ is a *conservative vector field*. Laurent's next step is to multiply (10.1) by a factor λ, and then to define a scalar function U, such that

$$dU = \lambda\left(p_1 dq_1 + p_2 dq_2 + \ldots + p_n dq_n\right). \tag{10.2}$$

Laurent actually uses Φ and μ rather than U and λ. I have chosen the latter for reasons that will become apparent, if they are not already, but for now we leave these terms uninterpreted. The reason for multiplying the differential by a factor will also become apparent. Because (10.1) is, by assumption, a perfect differential equation, there must be such a scalar function. Since we have (equivalently) assumed that **p** is a conservative vector field, **p** is proportional to the *gradient* of U:

$$p \propto \nabla U \tag{10.3}$$

where the gradient is the vector of the n partial derivatives of the scalar function U. The factor of proportionality is λ.

$$\lambda p = \nabla U \tag{10.4}$$

This is the integrability condition for the price function (the inverse demand function). It follows that the integral

$$\int_C \lambda p \, dq$$

is *path-independent*. That is to say, evaluating the integral will yield the same quantity regardless of what the agent's consumption-path through commodity space C was—regardless of what order he consumed his units of commodities in. Furthermore, by the gradient theorem,

$$\int_C \lambda \boldsymbol{p} \, dq = U\left(\boldsymbol{q}\right) - U\left(\boldsymbol{q}_0\right)$$

where $\mathbf{q}_0$ may be taken to be null, so that

$$\int_C \lambda \boldsymbol{p}\, dq = U(\boldsymbol{q}).$$

To return to Laurent's argument, it is then simple to show that U is a utility function. (10.2) above implies that

$$\frac{\partial U}{\partial q_1} = \lambda p_1; \frac{\partial U}{\partial q_2} = \lambda p_2; etc.$$

Marginal utility is thus proportional to price. This is exactly as it should be, if the consumer is maximizing his utility through his expenditure, given the law of diminishing marginal utility. And the factor of proportionality—the factor which Laurent introduced in (10.2) above—we can now see is a Lagrange multiplier. What Laurent is doing when he introduces this factor is using the Lagrangian method for solving the optimal consumption problem. This is the core point in my interpretation of Laurent's text. What he does not know, and what he inquires about towards the end of the letter, is what the economic interpretation of this factor is. We now know it to be the marginal utility of income, which is equal to the marginal-utility/price ratio of every commodity at optimal consumption.

Laurent then notes that the utility a consumer gets from having consumed a quantity of a commodity is found by integrating the marginal utilities

$$\int \frac{\partial U}{\partial q_1}; \int \frac{\partial U}{\partial q_2}; etc.$$

Here Laurent's chain of reasoning ends. We can complete it by defining the consumer's total utility from having consumed commodity-bundle **q**, *U(* **q** *)*, as the sum of the utilities he gets from having consumed each type of commodity in the quantity consumed. Total utility can be expressed as a line integral

$$U(\boldsymbol{q}) = \int_0^1 \sum_{i=1}^{n} U_i\left(q(t)\right) dq_i(t)$$

where $q(t)$, $0 \leq t \leq 1$, is the consumption path taken by the consumer. Given that (10.2) above is a perfect differential equation, this integral is path-independent, and so the total utility derived from consuming a bundle of commodities does not depend on the consumption path.

We have thus derived a total utility function for a consumer from his observed behavior. But—and here is the crucial point—we have only been able to do so by assuming that $p_1dq_1 + p_2dq_2 + \ldots + p_ndq_n$ is an exact differential. If we do not make this assumption, then $\lambda\mathbf{p}\, dq$ is *not* integrable. That is to say, the first integral given above

$$\int_C \lambda p\, dq$$

is *not* a function, but rather a *path-dependent functional*. When we evaluate it, the quantity we get will depend on which consumption-path has been taken. And if that is so, then defining a total utility function becomes *impossible*. Given that a consumer has consumed a certain bundle of commodities **q**, the question of what his utility from consuming that bundle is *has no unique answer*. The answer will depend on his consumption-path. If there is no utility function, there obviously is no such thing as maximizing the value of one's utility function. The assumption is foundational to GET. But as Laurent points out at the end of the passage, there is something troubling about it. It has no natural economic interpretation, and so it is unclear whether there is any motive for accepting it that is not *ad hoc*. This fact, as Mirowski documents, is well known among contemporary economists. The assumption is simply accepted as an odd but presumably harmless technical condition (Mirowski 1989, p. 370). What is not well known, and what we will see shortly, is that this assumption has significant negative consequences for the theory.

Now as I said, the interpretation of the argument presented by Laurent in this letter is a matter of some controversy. Laurent sent the same text to Vilifredo Pareto, and this section of the Laurent-Pareto correspondence has been discussed by J.S. Chipman (1976). Pareto denies that the scalar function on the left-hand side of (10.2) is a utility function—or, as Pareto would have called it, an "ophelimity" function—and argues that Laurent has misunderstood the concept of utility and the way utility functions are defined. Chipman describes Pareto as "proceed[ing] to spell out his objections [to Laurent's derivation of the utility function] very patiently and in utmost detail, in a respectful yet devastating fashion" (Chipman 1976, p. 43). Chipman goes on to speculate, however, that Pareto may have misinterpreted Laurent's derivation, and suggests that Laurent may have been looking for a way to define "an invariant cost-of-living index, i.e., an index expressing the minimum cost of a given level of well-being (utility) at current prices relative to the minimum cost of the same level of well-being at base-year prices, such index being invariant with respect to the level of well-being" (Chipman 1976, p. 44).

What is truly astonishing is that Chipman does not consider the possibility that Laurent meant his factor to be taken as a Lagrange multiplier, and that he was not only looking for, but had in fact found, a way to derive a path-independent utility function (something Pareto had failed to do, though he did not realize this at the time). It is astonishing for two reasons. The first is that Chipman identifies this as a possible interpretation of the factor, and indeed, as the interpretation which solves the problem of defining path-independent utility. The second is that he makes note of the fact that Laurent wrote a text book on calculus, in which the Lagrangian method is described in detail, and knew that Pareto had read and admired that textbook (Chipman 1976, pp. 43–44). But Chipman's very next remarks make it clear why he fails to even consider this interpretation of Laurent's argument. No economist managed to come up with this elegant solution to the problem of stating the integrability condition needed to define a utility function until Harold Hotelling in

1935 (Chipman 1976, p. 44). The first economist to solve the problem and state a correct condition was Eugenio Slutsky in 1915, by way of a far more complicated method (Chipman 1976, p. 44). Had Laurent's argument been interpreted in the way I have suggested, it "might have saved considerable work on the part of Walras, Pareto, and their contemporaries" (Chipman 1976, p. 44). But Chipman is apparently unable to believe that Laurent, despite being a master of the Lagrangian method himself and believing himself to be writing to others who understood it thoroughly, could have meant his argument to be taken this way, on the grounds that it could not have taken economists so long to recognize the ideas I am attributing to him. I find such an argument not at all convincing.

Chipman would likely respond that such an argument is undermined by the text of Laurent's letter. There are two issues, both of which may be dealt with quickly. The first is Laurent's use of the phrase "a standard measure of utility" for what is denoted by his factor μ (i.e. my factor λ). But I see no reason why Laurent's use of this phrase is inconsistent with the notion that he intended his factor as a Lagrange multiplier, and that the economic interpretation he was seeking for it was as the measure of the marginal utility of income, which at optimal consumption is equal to the marginal-utility/price ratio of every commodity consumed. The second is the somewhat tortured syntax of the sentence in which this phrase is used:

> If one accepts that there is a standard measure of utility, then one must also accept that expression (1) is integrable after having been multiplied by a factor μ [i.e. λ], if it is an exact differential.

My interpretation of this sentence is that Laurent is stating that if (a) we assume that (10.1) above is an exact differential equation, and (b) we accept that it is meaningful to introduce a Lagrange multiplier and apply it to (10.1)—which is what he means by accepting that there is a standard measure of utility—then, since (c) exact differentials are integrable, by (d) multiplying (10.1) by the Lagrange multiplier we will (e) be able to posit a scalar function Φ (i.e. U) which we will (f) be able to show is a utility function. Though this interpretation is certainly not the most natural reading of the sentence in isolation, I think it is the only one which is plausible given the broader context.

So much for the small controversy over Laurent. Now for the big one over Mirowski. The overall goal of Mirowski's work is to show that neoclassical economics, from which GET and thus all contemporary orthodox economic theory is descended, is fatally flawed. The heart of his argument is an attempt to demonstrate that the founders of neoclassical economics defined the central concepts of utility and expenditure so that they would be strictly analogous to the definitions of potential and kinetic energy, respectively, of classical mechanics. This analogy, Mirowski points out, is a bad one. Energy is conserved. The total energy of a closed system, its potential energy plus its kinetic energy, is a constant. But there is no analogous economic conservation law. This observation motivates Mirowski's interest in the integrability problem and the assumption required for its solution. The assumption which is needed to solve the integrability problem in economics has an analogue in physics: that there is a scalar function whose gradient equals the momentum vector

of a Hamiltonian system. The dynamics of the Hamiltonian system will then be integrable. The condition for there being such a function is that there be n conservation laws of the system, where the value of n depends on the number of dimensions of the system's phase space (McCauley 2000, p. 513). Since there are no conservation laws in economics, there is no natural economic interpretation of the assumption. So neoclassical economics defines its concepts based on faulty analogies with physics, rather than defining them in a way that would make them conducive to accurately representing actual economic phenomena. Present-day economists have forgotten the questionable origins of a number of the assumptions they take for granted, and so are modeling an economic world that resembles the world represented in nineteenth century mechanics more than it does any actual economy.

Mirowski's work, unsurprisingly, has been immensely controversial. Some of the negative reviews his work has received are irrelevant, focusing almost exclusively on the dismissive tone that characterizes much of his writing.[7] The most substantive negative review is that by the economist Hal Varian, who quite sensibly argues that the existence of points at which the analogy between economics and nineteenth century physics fails does not in itself show that there is anything wrong with economics, and these points should simply be taken as respects in which the concepts of the two disciplines are different. I agree with Varian's assessment of the force of Mirowski's argument. Although Varian does not make this point explicitly, the analogy of [potential energy : kinetic energy :: utility : expenditure] which is at the heart of Mirowski's argument is not as strict as he believes it is. In fact, potential energy is not even the appropriate physical analog of utility as utility is defined in neoclassical economics. As the physicist Joseph McCauley, who was inspired to examine the dynamics of GET after reading Mirowski's work, has pointed out, the concept of utility is actually analogous to the concept of action in mechanics—utility maps a path through commodity space onto a real number (McCauley 2000, p. 509). I cannot, however, share Varian's opinion that all is well with GET. As I said, I am not persuaded by Mirowski's argument that neoclassical economics is untenable because the integrabiliy condition required to define utility functions is adopted from mechanics and has no natural economic interpretation. But what McCauley has shown is that this assumption does create serious problems for GET. So it is to McCauley's work that we finally turn, to find the conclusions which undermine the article of faith with which this section started.

McCauley's argument is dense and highly technical. A detailed summary is beyond the scope of what is appropriate in the current context—the interested reader may consult his work directly. I will provide only a brief informal summary. What we are interested in is the movement over time of the price- and output-levels of a production economy whose participants consume and produce so as to maximize their total utility subject to their budget and production constraints. What we want to know is whether utility-maximizing behavior could result in the economy eventually reaching a stable position at which excess demand for all commodities has been reduced to zero—a stable equilibrium. Answering this question requires

[7] See, for example, (Hoover 1991).

modeling the economy as a dynamical system, evolving in time in a way determined by the disequilibrium forces that affect it. The key step in McCauley's argument is to show that the economy, when considered dynamically, can be modeled as a Hamiltonian system, with generalized coordinates of price and production corresponding to momentum and position (McCauley 2000, p. 515). Hamiltonian dynamics may be either non-integrable (and so path-dependent) or integrable (and so path-independent). Assuming the integrability condition discussed above gives us integrable, path-independent dynamics.

The motion of a Hamiltonian system may be either bounded or unbounded. The full phase space of the system represents all the combinations of coordinates—so in the case where the system is an economy, all the combinations of price-level and output-level—which the system can possibly reach consistent with some possible set of initial conditions—some possible set of consumer preferences, technological limits, and distribution of pre-trade endowments. The motion of the system is bounded for initial conditions from which it cannot reach every point in the full phase space, unbounded for initial conditions from which it can. What McCauley shows is that in the case of bounded motion, *all economic equilibria are unstable*. That is to say, even though all participants are acting so as to maximize their utilities subject to their constraints, the economy will *not* arrive and remain at an equilibrium, where price, production and consumption are such that excess demand is zero. Price and production levels will continue to change with time indefinitely, and we may see stable or unstable (i.e. non-chaotic or chaotic) oscillation, either with no approach to equilibrium, or with the system passing through equilibrium and then out of it again without the occurrence of a shock (McCauley 2000, pp. 516–517). The case of unbounded motion is not much better. The system will then have stable equilibria, but these will be hyperbolic. For our purposes, what is important about this fact is the extreme constraints it places on the possibility of reaching a stable equilibrium. The first is a pre-trade lump-sum endowment transfer, to place ourselves in a set of initial conditions consistent with unbounded motion and lying on a stable asymptote. The second is that the price-adjustment mechanism "is presumed to be infinitely precise and certainly cannot be subject to any noise" (McCauley 2000, p. 516).

This is unsurprising given the result achieved by Saari in the pure exchange case, which is that convergence on a stable equilibrium is only guaranteed given a price-adjustment mechanism that satisfies "an infinite information requirement… If there is an upper bound on the amount of information used in the adjustment process, i.e., if it relies solely on information about any fixed number of past periods and any fixed number of derivatives of the excess demand function, then there are cases in which the process fails to converge" (Ackerman 2002, p. 123). But the conclusion to be drawn from McCauley's work is a stronger one. In his production economy model, the *only* cases of convergence are cases of unbounded motion converging on hyperbolic equilibria. And McCauley is able to diagnose the source of the problem: it is the assumption which forces the dynamics to be integrable, the assumption needed to define total utility functions. As he repeats more than once, this "is not an equilibrium condition" (McCauley 2000, pp. 516–517). So the moral of the story is

this: given a production economy and a non-*tâtonnement* price-adjustment mechanism—in other words, given the real world—the only way for a free-market economy (one with no externally imposed constraints) made up of utility-maximizing agents to arrive at a stable equilibrium is by starting with an appropriate pre-trade lump-sum endowment transfer and making use of an infinitely precise price-adjustment mechanism. Even if the former were satisfiable, the latter is an impossibility. Since free-market economies have no inherent long-run tendency to arrive at stable equilibrium, the debate between New Keynsians and New Classicals is fundamentally misguided. And the predictive failures of general equilibrium models in general undercut the idea that models which do make these outlandish assumptions in order to generate equilibrium outcomes do a good enough job, for our practical purposes, of representing the workings of actual economies.[8]

The only possible conclusion is that a free-market environment cannot lead to a stable equilibrium in the real world, and thus cannot actually lead to an efficient allocation of resources. This conclusion, moreover, cannot be avoided by replacing the *homo economicus* agents of traditional economic models with a more realistic type of agent. It holds up, for example, against Herbert Gintis' recent demonstration of a fairly stable and efficient equilibrium arising from an agent-based model with a non-*tâtonnement* price-adjustment mechanism in which agent's actions are determined by a strategy of imitating those who are better-off than they are (Gintis 2007). The reason is that the exclusion of capital goods is essential to the success of Gintis' model (Bilancini and Petri 2008). Whatever degree of stability and efficiency real economies actually achieve is due to some other set of causes. And as we will see, the most likely candidate is a sound set of social, political and legal institutions, which both shape and participate in the market in a way that far exceeds not only the public-safety and contract-enforcement functions of a minimal State, but the limited policy interventions of New Keynesianism as well.

3 Evolutionary Economics

3.1 Revisiting Keynes: Money, Time and Uncertainty

We have looked at the problems that plague GET from one perspective, focusing on the assumptions it must make about the structure of markets, the production process, and the price-adjustment mechanism. But the path to a more promising approach to understanding the workings of real economies becomes clear when we turn our attention to a different set of assumptions, which concern the nature of the individuals who make up the economy. John Maynard Keynes' work, particularly his magisterial *The General Theory of Employment, Interest and Money*, can be viewed as an attempt to reintroduce three important elements of actual economic

[8] See note 6.

life into economic thought; three elements that he saw as wholly neglected by neoclassical economics (Keynes 1936/1964). These three elements are money, time, and uncertainty. Keynes, like the institutional economist John R. Commons, saw that for those who live in a monetary-exchange economy—an economy in which money is used as the medium of exchange—the very existence of money has an important effect on their preferences and their behavior. As the post-Keynesian macroeconomist Victoria Chick explains, "Money, as is well known, permits the separation of the act of selling goods from the act of purchasing them; that is, indirect exchange…Indirect exchange means a separation in time between actions involving real goods. The real value of a sales transaction, therefore, cannot be known for certain. In that sense, every transaction is a speculation…" (Chick 1983, p. 5). In this observation, the other two elements, time and uncertainty, are introduced, and the relation between the three becomes clear. Money matters because the existence of money introduces a temporal separation between selling one good (such as one's labor) and acquiring another (such as whatever one purchases with one's wages). But the passage of time between the two halves of an exchange of real goods matters (and thus money matters) because of our fundamental uncertainty about the future. From the perspective of the producer, as opposed to the wage-laborer/consumer, it is even clearer that of these three factors, uncertainty is the most significant. The existence of money serves to exacerbate the problems caused by uncertainty about the future. As Chick puts it:

> The time-consuming nature of production places upon producers the necessity to make decisions based on an *estimate*, a forecast, of the demand for their product: the goods must be placed on the market before people can buy them, and thus before demand can be known. The existence of money can enhance the difficulty of making that estimate, for when people save for future purchases, they do not need to make specific order even if they know what they will want and when. They can hold money instead… (Chick 1983, p. 5)

The basic decisions of all participants in an economy—how much to produce, what wage to work for—are based on expectations about the future: future demand, future costs, future prices. "These cannot be known for certain, but commitments must be made regardless" (Chick 1983, p. 11).

Keynes' goal was to understand the workings of an economy made up of participants engaged in monetary exchange, acting in time and uncertain about the future. His answer overturned Say's law, the cornerstone of the neoclassical theory of employment, which implied that it was impossible for the economy to arrive at a stable level of employment and outcome short of full employment. Keynes' key observation was that there are circumstances, such as the aftermath of the bursting of an asset bubble, in which individuals have a strong desire to hold their assets in liquid form—a strong liquidity preference—since they see this as a way of protecting themselves against additional but unforeseeable calamities in the future. A strong liquidity preference throughout the community translates into a high rate of interest, a high premium for parting with some of one's liquid assets. A high rate of interest discourages investment by firms, and so depresses aggregate demand, one part of which is the sum of payments made by firms to other firms for capital goods, in excess of the aggregate user cost (which is what firms sacrifice in the value of

their capital goods for using them to produce goods). In such an environment, firms will not hire additional employees at the going wage, even if they are operating below their current productive capacity. This is not because the going wage is too high (as the neoclassical theory would claim), but because the rate of interest is too high. Hiring additional employees increases aggregate income, and with it aggregate consumption, since increased income always brings with it increased consumption. But—and here is another of Keynes' key psychological observations—spending on consumption does not increase as quickly as income. We have a diminishing marginal propensity to consume. The more we earn, the greater the percentage of what we earn that we save. Increased employment, then, boosts aggregate demand (by boosting aggregate consumption), but given the propensity to consume, the increase in employment is only worthwhile from the firms' perspective if the amount saved is borrowed and invested. That investment provides the second leg of the boost in aggregate demand, which is needed to make the additional employment worthwhile to the firms. This increased investment is what does not occur when interest rates are driven up by a strong liquidity preference—which is to say, by fear of what an uncertain future might hold. In such a scenario, cutting wages will do nothing but diminish aggregate demand even further, by diminishing aggregate income. This leads to further reductions in production, and so less employment. It is a recipe for economic depression.

Keynes' diagnosis of this situation is the source of his macroeconomic policy suggestions. The first route back to full employment is through the Federal Reserve, which can endeavor to drive the rate of interest down by buying bonds. This increases the supply of money, making money less scarce. It is possible, however, that the strength of the liquidity preference will grow faster than the money supply, so that the Fed's action has no effect on interest rates. This is the famous Keynesian liquidity trap. Keynes discusses two ways of getting out of it. One is for the government to boost aggregate demand directly through deficit spending. The other is through progressive redistribution, since the marginal propensity to consume of the poor is higher than that of the rich. Keynes thus provides an argument for preventing extreme wealth inequality that appeals to the economic value of efficiency, rather than to the value of equality itself. The latter two options are the only ones available in a so-called "balance-sheet recession," such as the one that afflicted Japan during the "lost decade" of the 1990s (Koo 2009). The additional component of that scenario is that firms find themselves with assets that are worth less than their liabilities, and focus on paying down debt rather than investing even after monetary policy has driven the interest rate down to zero.[9]

Since real-world economies are populated by agents engaged in monetary exchange, acting in time and uncertain about the future, it might seem obvious that this is what economic theory should strive to understand. Nonetheless, orthodox economics after Keynes has come as close to abandoning this goal as it possibly

[9] Strictly speaking, there may still be a role for monetary policy after the interest rate has reached zero. Paul Krugman famously argued that the central bank could engage in "managed inflation," and turn the real interest rate negative (Krugman 1998).

could, without ignoring Keynes altogether. The reason for this lies in a problem with Keynes' methodological approach, and understanding this problem is the key to recognizing the direction that economic theory should take following the failure of GET. Keynes' approach to macroeconomics is based on a set of hypotheses about relationships between aggregate variables—quantities like aggregate consumption, aggregate investment, and aggregate income. The controversy surrounding Keynes' work has centered on the question of whether these hypotheses were consistent with any acceptable set of assumptions about the behavior of the individuals whose actions make up these aggregates. The way this is usually put is by asking whether Keynes' macroeconomic theory has any acceptable microfoundations. This is the heart of the so-called "Lucas critique," named after the economist Robert Lucas, of Keynes' approach to macroeconomics. Since those who were asking the question were nearly unanimously of the view that the only acceptable set of assumptions about the behavior of individual economic agents is that posited by neoclassical microeconomic theory, it should be unsurprising that the consensus answer was "No."

Three important developments came out of this controversy. The first was a change in the neoclassical paradigm which was meant to accommodate Keynes' unassailable observation of the importance of uncertainty about the future for modeling the behavior of economic agents, but in a way which would have as small an impact as possible on the conclusions of traditional neoclassical economics. This was accomplished in the form of the theory of Rational Expectations, developed primarily by Lucas, according to which the predictions of all future values of all economically significant variables by economic agents contain only random errors, such that correct predictions can be recovered by averaging these expectations over a long enough time. The second development was the formulation of the representative agent assumption, which requires the assumption that individual consumer preferences are homothetic, in order to avoid the S-M-D result. This conflicts with Keynes' central observation that the propensity to consume increases more slowly than income.[10] The third development followed on the first two. It was the acceptance of Rational Expectations theory and the representative agent assumption by orthodox economists of Keynesian persuasion, and an attempt to formulate a version of Keynesian Macroeconomics consistent with this underlying microeconomic theory. It is this effort which gave birth to New Keynesian macroeconomics.

Given the failure of GET, the unquestionable validity of Keynes' goal, and the need to model macroeconomic phenomena in a way that intelligibly connects them to microeconomic behavior, the appropriate future direction for economic theory is clear. What is needed is a different approach to understanding and modeling the behavior of individual economic agents, one that breaks with the hopelessly implausible assumptions of neoclassical/Rational Expectations theory, and reintroduces the fundamentally important role played by uncertainty about the future in determining economic behavior. The version of economic theory that fits the bill is evolutionary economics.

[10] See note 4.

3.2 Evolutionary Economics and the New Microfoundations

Evolutionary economics is based on the evolutionary game theory of John Maynard Smith (Smith 1982). It models economic agents as boundedly rational. That is to say, agents are assumed to have preference-orders, and thus utility functions which they are interested in maximizing, but they are not assumed to be capable of straightforward maximizing behavior. They are not assumed to have "rational expectations" of the future, or to be capable of determining an optimal course of action in every choice situation. Instead, they seek to satisfy their preferences through the adoption of strategies, and rules for the revision of strategies over time. Their preferences may change over time. And they need not be purely self-interested in any narrow sense; rather, they may be disposed to cooperate, and to follow public rules and policies, even when doing so involves some sacrifice. The normal state of an evolutionary market is not assumed to be one of stable equilibrium. Evolutionary models recognize that economies are dynamic, non-linear, and often out of equilibrium.

In order for the idea of evolutionary economics to be more than metaphorical, we need to understand the process of evolution outside of a biological context. Eric Beinhocker has done an admirable job setting out the necessary conditions for, and components of, a process of evolution in a suitably general way, and mapping these onto an economic environment (Beinhocker 2006). He identifies six necessary conditions for evolution to take place: (1) a design space, that is, a set of all possible designs a type of entity could have; (2) schemata, made up of building blocks, which code designs; (3) interactors made up of modules which instantiate the design coded by the building blocks of a schema; (4) schema-readers, who render schemata into interactors; (5) an environment into which interactors are rendered, which places constraints on them; and (6) a function that measures the fitness of interactors (Beinhocker 2006, pp. 213–214). We then conceive of evolution as a search algorithm, which searches the design space for designs that are fit, given the environmental constraints. This is a recursive process, and one in which variation is introduced into schemata over time. The interactors rendered by active schemata form a population. These interactors replicate according to their fitness, with the result that modules contributing to fitness become more frequent over time (Beinhocker 2006, p. 214). The results of this process are the emergence of order from randomness, novel adaptations to changes in the environment, the discovery of fit designs, and growth in the amount of resources devoted to fit designs (Beinhocker 2006, pp. 214–215).

This very abstract model of evolutionary selection makes a precise statement of the components and structure of an evolutionary economic model possible. There are three design spaces—the space of all possible business designs, the space of all possible physical technologies, and the space of all possible systems of human organization. The schema that codes for a particular business design-technology-organization combination is a business plan, and the schema-reader is a management team. The interactors are the firms which instantiate the active business plans. Their

environment is the marketplace, and the business plans include the strategies which guide the actions of the firms in the marketplace as they compete for scarce resources (capital goods, raw inputs, and consumer dollars). This competition is never assumed to be perfect in the neoclassical sense—all of the assumptions of neoclassical prefect competition may be violated in an evolutionary model of the marketplace.[11] Rather, the level of competitiveness in an evolutionary market can be judged far more intuitively according to the number of firms in a market and how evenly marketshare is distributed among them. Firms are composed of modules. A module is anything, including most obviously a piece of technology, a system of organization, or a strategy used by the firm, which has provided or could in the future provide a basis for differential selection among firms in the marketplace (Beinhocker 2006, p. 281). Fit firms are the ones who replicate—in the next period, they open new divisions or branches in new locations, enter additional markets, etc.—and unfit firms eventually close. Some smaller niche firms will simply stay in business without growing (think of a firm that does this as replacing itself with a single exact replica at the end of each period). Modules are the units of selection. From one production period to the next, the modules that enhance fitness will become more frequent. Consumers are likewise boundedly rational, and are modeled as having preferences which change over time.

Evolutionary economic theory uses the methodology of agent-based computational economic modeling. Such models begin with populations of distinct avatars, each representing an individual firm, consumer, household, bank, or other economic participant. The behavior of these avatars is guided by their preferences, their limited information about the marketplace, their biases, their strategies for interaction, and rules for revising their strategies over time and updating their store of information. These elements differ between each avatar, and are based on empirical behavioral data. The models then simulate the interactions between the agents represented by the avatars and allow macroeconomic patterns to emerge from these individual interactions. They can be tested by running simulations with initial conditions based on historical data, and observing how closely those simulations then recreate macro-level patterns that match those captured in the macroeconomic data of that era. This process is called empirically validating the model. The empirically best validated models can then be used to predict the likely effects of policy changes, by running new simulations with key environmental variables altered.

These are the alternative microfoundations that were required by, but unavailable to, Keynes. An agent-based computational model need make no brute assumptions about the relationships between macroeconomic variables. Hypotheses about the relationships between macroeconomic variables can be tested in the model, as simulations reveal the conditions under which the hypothesized macro-level patterns do and do not emerge. Of course, this is all easier said than done, and the validity of the results of any such test will depend on the extent to which the model accurately represents the behavior of real economic agents. There are many questions regarding what information should be used in designing avatars, what may be safely left

[11] See note 2.

out, and how the empirical behavioral data used to design avatars should be collected. These questions do not have obvious answers. Agent-based macroeconomic modeling is still young. But there are already a number of robust, empirically well-validated results which are extremely encouraging to followers of Keynes, as we will see in the next section.

3.3 The Evolutionary Free Market and Its Limits

GET provided one argument to justify allowing markets to operate free of interference: free markets reach equilibrium—and thus achieve stability in price- and output-levels—and reaching equilibrium entails allocating resources efficiently. GET, as we have seen, is a failed program. Evolutionary economics offers its own very different justification for the same policy. As Beinhocker puts it:

> Markets provide incentives for the deductive-tinkering process of differentiation. They then critically provide a fitness function and selection process that represents the broad needs of the population (and not just the needs of a few Big Men). Finally, they provide a means of shifting resources toward fit modules and away from unfit ones, thus amplifying the fit modules' influence. In short, the reason that markets work so well comes down to what evolutionary theorists refer to as Orgel's Second Rule (named after biochemist Leslie Orgel), which says, "Evolution is cleverer than you are." Even a highly rational, intelligent, benevolent Big Man would not be able to beat an evolutionary algorithm in finding peaks in the economic fitness landscape. Markets win over command and control, not because of their efficiency at resource allocation in equilibrium, but because of their effectiveness at innovation in disequilibrium. (Beinhocker 2006, p. 294)

What is interesting about this argument is not that it succeeds where the neoclassical argument failed—for as we will see in a moment, it does not succeed. Rather, the reasons why it does not succeed point the way to the revival of traditional Keynesianism and "Old" Institutionalism in economic theory—the Institutionalism of Thorstein Veblen, John R. Common, and J. K. Galbraith—by revealing the points at which the new evolutionary economics needs to be supplemented by, and integrated with, these older theories. Old Institutionalism was in fact the first branch of economic theory to emphasize the importance of the fact that economic systems evolve.[12] In what follows, I will assume the contemporary evolutionary-institutional economist Geoffrey Hodgson's definition of "institutions" as "durable systems of established and embedded social rules that structure social interactions. Language, money, law … firms (and other organizations) are all institutions" (Hodgson 2002, p. 113).

[12] I refer to this school of economic theory as *Old* Institutionalism to distinguish it from the *New* Institutional Economics which grew out of the work of the economist Ronald Coase. New Institutionalism is an attempt to place Old Institutionalism on neoclassical microfoundations. It is a parallel effort to New Keynesianism, and just as barren. The earliest argument that evolution, rather than classical mechanics, was the appropriate natural-scientific role-model for economics was made in a seminal essay by Thorstein Veblen (Veblen 1898).

The considerations that undermine the evolutionary argument for free markets can be grouped into two categories: problems inherent to evolutionary markets, and limitations of the evolutionary model itself. Both types of difficulties point to the need for strong public institutions to both shape and participate in the market. Under the first heading belong the problems of the volatility of the business cycle, the human cost of a pure market-based society, and the wastefulness of competition for positional goods (i.e. commodities whose sole or primary purpose is to confer social status). These difficulties indicate the need for public institutions prepared to take Keynesian policy action. Under the second (which we shall examine in the next section) fall the precariousness of the conditions of competition, the ability of large firms to change the environment in which they operate, and the uneasy relationship between business and democracy. These point toward the necessity of going beyond Keynes while preserving his insights, by integrating them into a revival of the Old Institutionalist school of economic theory, which must in turn be integrated with the new evolutionary school.

A Keynesian evolutionary perspective on market activity (i.e. an evolutionary model of the market that represents agents as possessing the psychological propensities discussed by Keynes) puts us in a good position to understand the causes of the business cycle—the recurring historical pattern of prosperity, recession and recovery.[13] Examining this phenomenon makes clear the first limitation of the evolutionary free market. Some cycles are caused by exogenous events, positive or negative shocks to the economy like natural disasters and major technological innovations. Some, on the other hand, are caused by the collective behavior of economic participants, not in response to a shock, but simply as an expression of their own psychological propensities, or natural physical and cognitive limitations. Within these broad categories, different types of business cycles can be distinguished based on both their periodicity and their specific cause. Traditional business cycle theory, as systematized by Joseph Schumpeter, distinguishes three types of cycle (Schumpeter 1961). Substantial evidence for the existence of cycles of these types has recently been acquired through spectral analysis of global data on the GDP of industrialized nations from 1870 to 2007 (Korotayev and Tsirel 2010).

The shortest cycle is the 3–4 year "Kitchin wave," named for the economist Joseph Kitchin (Kitchin 1923). These cycles are caused by price, wage, and information stickiness, and so can be explained by New Keynesian models, as well as by evolutionary ones. The transition from prosperity to recession is caused by a drop-off in aggregate demand due to the economy reaching its expansionary limit after a period of growth. Firms that do not or cannot respond quickly enough to this—because the necessary information does not reach them quickly enough, or they are already committed to a certain level of capital investment based on expectations of future demand that now prove overly optimistic—end up making deeper cuts to their labor force and operating further below their productive capacity than they would have had to had they responded more quickly. But so long as prices do

[13] See, in particular, (LeBaron and Tesfatsion 2008; Dosi et al. 2006; Dosi et al. 2010; Dosi et al. 2013).

eventually decrease sufficiently, aggregate demand begins to recover and increased employment follows. The rate of interest is not an important factor in the context of the Kitchin wave, because of its short duration. Aggregate demand from capital investment can be treated as fixed over the period.

Next in periodic length is the 7–9 year "Juglar wave," named for the nineteenth century French economist Clement Juglar (Juglar 1862). The period of a Juglar wave is long enough that firms make new decisions regarding levels of capital investment for future production within it. Within a Juglar wave, the transition from expansion and prosperity to recession can have either an exogenous or an endogenous cause. Both the traditional Keynesian analysis discussed above and New Keynesian analysis focus on Juglar recessions with exogenous causes—negative shocks to an economy which depress aggregate demand. Sophisticated New Keynesian models can employ additional market imperfections—such as the existence of monopoly or oligopoly competition, or stickiness in the credit market as well as in the real economy—to explain such cases in a way consistent with their assumptions of rational expectations and homothetic consumer preferences. Of course, these assumptions make the explanation less realistic than that given by a Keynesian evolutionary model.

But these recessions can also have endogenous causes, arising from the operation of what Keynes famously called "animal spirits":

> Even apart from the instability due to speculation, there is the instability due to the characteristic of human nature that a large proportion of our positive activities depend on spontaneous optimism rather than mathematical expectations, whether moral or hedonistic or economic. Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as the result of animal spirits—a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities. (Keynes 1936/1964, p. 161)

These endogenous Juglar recessions are triggered by bursting asset bubbles. The driving up of prices that is observed during the inflation of an asset bubble is an illustration of what evolutionary and institutional economists call the principle of circular and cumulative causation. This is an idea developed by the economist Gunnar Myrdal, in order to explain the disequilibrium dynamics of economic systems (Myrdal 1957). In essence, it is the observation that price-levels are affected by positive feedback mechanisms.[14] The wildly optimistic expectations (the animal spirits) of some with respect to, for instance, the direction of the prices of internet stocks (as in the Dotcom bubble of the late 1990s), need not result in an equal and opposite reaction on the part of others who recognize the temporary nature of that unjustified optimism, and see the upward pressure it exerts on prices as creating an opportunity to make a profit in the near future by short selling. Such a reaction is what would be required in order to return prices to a stable equilibrium. Instead, the

[14] The existence of positive feedback mechanisms within the market is in line with an evolutionary model of the economy. For a survey of the role of positive feedback mechanisms in evolutionary biology, see (Crespi 2004).

optimism of some breeds further optimism in others not wanting to be left behind, and so on, until prices are far out of line with any valuation based on the fundamentals of the companies in question. When reality catches up with expectations—when, for example, many highly valued companies, which have been borrowing furiously to fund rapid expansion in order to reach more customers, must finally disclose how unmanageably wide the gaps between their debts and their earnings have become—a bust, a bursting of the asset bubble, follows. Since the sub-prime mortgage crisis of 2007, there has been a renewal of interest in credit/debt cycles, and particularly in the post-Keynesian economist Hyman Minsky's "financial instability thesis," which claims that the crucial stage in the spread of unrealistically optimistic expectations is the point at which they begin to influence the behavior of banks, who aggressively compete to extend credit in the belief that future returns on investment will continue to increase.[15] The subsequent bust then inflicts damage on the institutions that constitute the financial system itself along with firms and households, resulting in a reduction in credit (due to an increase in liquidity preference among financial institutions) which makes loans unaffordable even for firms and individuals who remain financially sound, and further reductions in aggregate demand.[16] Keynesian evolutionary models are much better equipped to explain these recessions than models that assume rational expectations (Dosi et al. 2013).

The final type of business cycle is the long "Kondratiev wave" (named for Russian economist Nikolai Kondratiev), with a period of approximately 50 years (Kondratiev 1925). The most influential explanation for these cycles is Schumpeter's "creative destruction" hypothesis (Schumpeter 1961). A Kondratiev recession results from a real economic development, such as a new technological innovation, which disrupts the business plans of a sufficiently large number of firms, and results in the sudden obsolescence of their capital investments and in the skills of their employees. The classic example is that of the Luddites in nineteenth century England—the textile artisans whose skills and tools became obsolescent and whose livelihood disappeared with the invention of the power loom. The long-term effect of such innovation is strongly positive: a return to full employment eventually follows, as the market realigns and new types of jobs are created in the environment dominated by the new technology; and this new technology, once integrated, makes possible new highs in levels of productivity and wealth-creation. But even though innovation does not lead to long-term reductions in the employment level (this is the so-called "Luddite fallacy"), we must remember the lesson from Keynes: there is no shortage of opportunities along the road to recovery for the economy to get stuck at a stable level of underemployment and underproduction. The adoption of new

[15] See HP Minsky (1986/2008) *Stabilizing and Unstable Economy* (New York: McGraw Hill). For a recent agent-based model with Minskyan credit/debt dynamics, see (Dosi et al. 2013).

[16] The reader interested in the sub-prime mortgage crisis will find many books to choose from. For the best narrative of the crisis itself and its aftermath, see (Ferguson 2012). For the most thorough diagnosis of the major economic transformations which made the crisis possible, and which have made the recovery so sluggish—including the rise of corporatism, oligopoly power, and the financialization of capitalism—see (Foster and McChesney 2012; Tabb 2012).

technologies often requires a great deal of upfront investment, and so the inducement to invest (and thus the rate of interest) may play an important role in the process of recovering from an industrial bust, at least if the innovation is sufficiently disruptive to have a significant impact on aggregate employment, and thus trigger a widespread increase in liquidity preference. And the period in between bust and recovery can be a long and painful one for many.

New Keynesian economics represents disruptive technological innovation as a random shock, a positive analog of a natural disaster. Evolutionary Keynesian models, on the other hand, allow us to explicitly model the search process which results in disruptive technological innovation. This is an illustration of what Beinhocker identifies as the primary benefit of the market from an evolutionary perspective: it is unbeatable as a source of innovation. And innovation is undoubtedly a social good, leading as it does to heightened future levels of productivity and wealth-creation. But Beinhocker's argument is defective insofar as his focus on the impersonal firm as the interactor in the evolutionary model screens off the real human cost incurred during a Kondratiev recession. There is also much to be lost and nothing to be gained by a failure of the State to boost aggregate demand in a Juglar recession (whether endogenous or exogenous), or to work to reduce information stickiness (à la Greenwald and Stiglitz) in a Kitchin recession. The volatility of the very business cycles which evolutionary models help us understand undermines the argument for a free evolutionary market. The social acceptability of a dynamic, non-linear market, with no natural tendency to stable and efficient equilibrium, is at least going to require the existence of a suite of redistributive social assistance programs to soften the blow of a bust, and a set of public institutions with sufficiently powerful monetary and fiscal policy tools to jump-start a stalled recovery. Endogenous Juglar cycles add nothing to the economy—the bursting of an asset bubble typically more than erases the gains in wealth accumulated during its inflation. Preventing, or at least mitigating, endogenous Juglar recessions requires public institutions powerful enough to act before they occur. This means smoothing out these cycles by enacting borrowing rules (such as sufficiently conservative debt/equity ratio requirements), and trading regulations (such as the requirement in Title VII of the Dodd-Frank Act that trades in derivatives be made through financial clearing houses); and making countercyclical monetary policy decisions to cool the wild optimism that inflates a bubble.[17]

Although the innovation an evolutionary market delivers is undoubtedly a social good, the stability and efficiency it does not provide are social goods as well. The fact that an evolutionary market does not deliver them on its own is not a reason for giving up on trying to achieve them, at least to a fair approximation. The features of individual psychology stressed by Keynes, which evolutionary models allow us to represent at a micro-level, and which give rise to the macroeconomic phenomena which Keynes himself examined, necessitate intervention in the market by public

[17] See Title VII – Wall Street Transparency and Accountability, The Dodd-Frank Wall Street Reform and Consumer Protection Act (124 Stat. 1376–2223). For a thorough argument in favor of requiring banks to maintain lower debt/equity ratios, see (Admati and Hellwig 2012).

institutions employing Keynesian policy tools. In the United States, the savings and loan crisis of the late 1980s, which followed a wave of financial deregulation, was the first instance of mass financial institution failure since the Great Depression—a period of 50 years.[18] In 140 years of American financial history prior to the New Deal, the country experienced a bank panic *every 10–20 years*. The loss of wealth and diminished efficiency, as well as the cost in human suffering, incurred during American recessions between 1937 and 2007 have also been far less severe than in the recessions of the pre-Depression era (Zarnowitz 1996). Financial regulation, counter-cyclical monetary policy and the provision of social assistance are essential to maintaining some degree of stability in the face of the fluctuations of an evolutionary market.

And although we would surely be willing to make some sort of trade-off between innovation, on the one hand, and stability and efficiency on the other, recent research conducted by the Italian economists Giovanni Dosi, Giorgio Fagiolo and Andrea Roventini indicates that no such trade-off is necessary (Dosi et al. 2010). Their agent-based macroeconomic model has revealed profound complementarities between factors influencing aggregate demand and drivers of technological change that affect both "'short-run' fluctuations and long-term growth patterns…[and] a corresponding complementarity between 'Keynesian' and 'Schumpeterian' policies in sustaining long-run growth paths characterized by milder fluctuations and relatively lower unemployment levels" (Dosi et al. 2010, p. 1748). The smoothing of business-cycles that Keynesian policy action makes possible, in other words, creates an environment which is even more conducive to technological innovation than is a free market. One important reason for this is that innovation requires capital investment and research—business activities which firms do not engage in when aggregate demand suddenly drops off at the beginning of a crisis. Policies that make crises less frequent and less severe thus foster innovation. The relationship between strong, independent public institutions and innovation is one we will return to in the next section.

In addition to the waste and destruction of recessions caused by burst asset bubbles, evolutionary markets foster wasteful competition for purely positional goods.[19] Positional goods are those whose value to their possessors is determined by the relative quantity in which they are held: satisfaction from possessing such goods derives from the fact that others have less of them. They are consumed conspicuously (to use Veblen's term) as a public indication of one's socio-economic status. The consumption of positional goods thus imposes externalities on others: the mere fact that consumer A now possesses more of some positional good than he did will entail that the value of consumer B's share is less valuable (in his own eyes and the eyes of others).[20] Competition to consume positional goods, in other words, has the structure

[18] For an outstanding study of the crisis and its causes, see (Black 2005).

[19] The term was coined by Fred Hirsch (Hirsch 1976).

[20] This is one of the primary explanations for the failure of happiness, in the sense of subjective satisfaction with one's lot in life, to increase with increases in income past a fairly modest level (Scitovsky 1976, ch. 7).

of an arms race (Frank 2011, ch. 5). And as such, it makes economic growth into a zero-sum affair: it becomes impossible for everyone to get ahead (Hirsch 1976, ch. 1). When consumption of positional goods begins to crowd out consumption of non-positional goods, economic growth no longer brings with it Pareto-improvements in the allocation of resources; the individual pursuit of advancement no longer translates into the advancement of society as a whole. Consumption which focuses on them channels demand away from the use of productive resources—resources which could be used to create non-positional goods. Since the consumption of non-positional goods does not itself impose negative externalities, any shift in consumption away from positional goods to non-positional ones involves a gain in efficiency: the consumers of non-positional goods become better-off without thereby making anyone else worse-off. Likewise for a shift away from consuming positional goods to providing funds for investment in productive capital, which can be used to create non-positional goods.

This sort of wasteful competition is another problem with the evolutionary market which can be ameliorated through the action of strong socio-political institutions. The economist Fred Hirsch concludes his classic study of the social harms incurred through increasingly stiff competition for positional goods with the recommendation that we use the tools of public policy

> to reduce the costs incurred by the individual, in responding to his own instincts, to orient his behavior to a social need…This approach…suggests a broad guideline for policy. It is to reduce the incidental benefits from positional precedence. The operational objective should be to pare down the contestants to those who most value the benefits that cannot be obtained in other ways. (Hirsch 1976, pp. 180, 183)

Hirsch focuses on the most important variety of positional precedence, positional jobs:

> These are the jobs at or near the pinnacle of professions and within businesses…As long as the nonfinancial attractions of positional jobs are strong, the salaries attached to them can be regarded as incidental benefits. Money can be earned elsewhere; the attractions of the job can be gained only from doing it. A reduction in the monetary attraction can be expected to reduce total demand for such jobs by shedding potential applicants for whom the pay advantage is dominant…The means of such a reduction could take a variety of forms. One possibility is a payroll tax related to the size of differentials in pay within the firm, combined with direct action by government and other public sector employers to reduce differentials applying to executive and high professional positions. (Hirsch 1976, pp. 183–184)[21]

Hirsch, like Keynes, is thus able to provide a compelling argument for policies that prevent wealth inequality from becoming too extreme by appealing to the efficiency-enhancing effects of those policies, rather than to the value of equality directly.

[21] This seemed like extraordinarily prescient advice in 2013, 6 years after the beginning of the Great Recession and four years after the recovery officially began, as the employment level in the US remained below its pre-2007 levels, and average CEO pay was 273 times greater than average worker pay (See The Center on Budget and Policy Priorities 2012; and The Economic Policy Institute 2012).

The Keynesian efficiency-based argument for limiting wealth inequality has in fact been reinforced and expanded by the most recent work in agent-based evolutionary Keynesian analysis. Dosi, Fagiolo and Roventini, along with Mauro Napoletano, have used their agent-based model to show that "different income distribution regimes heavily affect macroeconomic performance: more unequal economies are exposed to more severe business cycles fluctuations, higher unemployment rates, and higher probability of crises. On the policy side, fiscal policies do not only dampen business cycles, reduce unemployment and the likelihood of experiencing a huge crisis. In some circumstances they also affect long-term growth" (Dosi et al. 2013, p. 1598). The Keynesian policies that yield these benefits of stability and efficiency are the same ones that limit wealth inequality. The effect on long-term growth is precisely the positive effect that Keynesian policies have on innovation. In the next section, we will see some additional reasons for complementarity between innovation and institutional policies that work to limit wealth inequality.

The conclusions of Dosi, Fagiolo, Roventini and Napoletano are nicely complemented by the econometric research of Jonathan Ostry and Andrew Berg of the International Monetary Fund. Working with four decades worth of data from over 150 countries, they reach three conclusions regarding the relationship between wealth inequality, redistribution, and economic growth:

> First, *more unequal societies tend to redistribute more…*
>
> Second, *lower net inequality seems to drive faster and more durable growth for a given level of redistribution…*
>
> Third, *redistribution appears generally benign in its impact on growth; only in extreme cases is there some evidence that it may have direct negative effects on growth.* Thus the combined direct and indirect effects of redistribution—including the growth effects of the resulting lower inequality—are, on average, pro-growth. (Ostry et al. 2014, pp. 6–7)

Moreover, "there is a strong negative relation between the level of net inequality and growth in income per capita over the subsequent period…and there is a weak (if anything, *positive*) relationship between redistribution and subsequent growth" (Ostry et al. 2014, p. 16; see also Berg and Ostry 2011; and Berg et al. 2012). Since lower inequality benefits growth *and* redistribution (at least of a non-extreme sort) has an at least neutral and perhaps weakly positive effect on growth, the attempt to achieve lower inequality (and thus greater, more durable growth) through the use of a set of policy and regulatory tools which includes policies that redistribute wealth need not have a negative impact, and may even have a positive impact, on growth *even in the medium term during the transition from greater to lower inequality*. Furthermore, since more equal countries redistribute *less*, greater equality, once achieved through policies that include redistribution, can be sustained without continuing to redistribute on as high a level as was necessary during the transition from greater inequality to greater equality.

The IMF's conclusions are echoed, and even amplified, by the findings presented in a recent report of the OECD:

> [I]ncome inequality has a negative and statistically significant impact on medium-term growth…The biggest factor for the impact of inequality on growth is the gap between lower income households and the rest of the population. The negative effect is not just for the

> poorest income decile but all of those in the bottom four deciles of the income distribution…[R]edistribution per se does not lower economic growth. (OECD Directorate for Employment, Labour and Social Affairs 2014, p. 2)

The OECD report, moreover, offers a clear explanation for this phenomenon: "The evidence is strongly in favour of one particular theory for how inequality affects growth: by hindering human capital accumulation income inequality undermines education opportunities for disadvantaged individuals, lowering social mobility and hampering skills development" (OECD Directorate for Employment, Labour and Social Affairs 2014, p. 3). The old argument that using redistribution to achieve greater equality will create a culture of dependence that will in turn necessitate the indefinite continuation of high levels of redistribution, is not borne out by the evidence. Redistribution is a ladder to greater equality which, if it cannot be thrown away once that greater equality is achieved, is one which we can put away until a time of crisis makes more aggressive Keynesian policies necessary once again. The message that is emerging from our Keynesian evolutionary perspective on the market is that the traditional market values themselves—stability, efficiency, productivity and growth, and innovation—flourish in the same environment as equality.

4 Reviving Old Institutionalism: Limitations of the Evolutionary Model

Our discussion of the evolutionary model of the market has led to an appreciation of the importance of public institutions capable of carrying out the sorts of policy actions advocated by Keynes, while the agent-based computational modeling employed by the evolutionary approach has emerged as a promising route to Keynesian microfoundations. But Keynes' revolution was an incomplete one (Galbraith 1973, p. 342). There are limitations on the representational adequacy of the Keynesian evolutionary model which his work does not help us to see. And so, while preserving Keynesian insights, we must go beyond Keynes and combine an evolutionary understanding of the micro-level with an institutionalist understanding of the macro-level, and of the interaction between the two.

Because real-world markets lack all the features of neoclassical perfectly competitive markets, the result of the competitive process in an evolutionary marketplace is the emergence of a few big winners, a large number of hangers-on, and an even greater number of losers—a small number of oligopolistic competitiors, alongside larger groups of small firms in more (but far from perfectly) competitive markets. From an empirical standpoint, there is very little doubt that we are living in an age of highly concentrated global oligopoly power. In the past few years, the first global studies of the structure of trans-national corporate ownership and governance have been conducted. Stefania Vitali, James Glattfelder, and Stefano Battiston have studied the network of direct and indirect ownership among over 43,000 transnational corporations, with the goal of stratifying them according to the level of

control each exerts over the global corporate network, where control over a firm is understood in terms of "the chances of seeing one's own interest prevailing in the business strategy of the firm" (Vitali, Glattfelder and Battiston 2011). The result of their study is that "transnational corporations form a giant bow-tie structure and that a large portion of control flows to a small tightly-knit core of financial institutions. This core can be seen as an economic 'super-entity'…" made up of roughly 700 business and financial entities which possess "80 % of the control over the values of all TNCs [trans-national corporations]" (Vitali, Glattfelder and Battiston 2011).

That short-term oligopolies should arise through the process of competition is no surprise from an evolutionary economic perspective. What fit firms do, after all, is replicate (that is to say, they grow). But the emergence of a stable oligopoly that spans virtually all industries and markets is a mystery from this point of view. The evolutionary model assumes that firms operate in a constantly changing environment. This would make stable business strategies short-lived. Continuous adaptation and innovation should be necessary, and the stable domination of the global marketplace by a core group of 500–1000 enormous corporate entities should be an exceedingly unlikely state of affairs. This observation brings us to the major blindspot of the evolutionary approach, the crucial fact which the evolutionary model does not represent and which Institutionalism brings into focus. *The large trans-national firm has a profound ability to change the environment in which it finds itself to suit its needs and goals, rather than changing itself to adapt to an environment outside its control.* As J.K. Glabraith puts it:

> Over the last hundred years numerous economic tasks have come to be performed by organizations – by industrial corporations, electric utilities, airlines, merchandising chains, banks, television networks, public bureaucracies. Some of these organizations are very largeas few would doubt, they have power, which is to say they can command the efforts of individuals and the state. They command these, most will agree, for their own purposes, these being the purposes of those who participate through membership or ownership in the enterprise…Partly the economic system serves the individual. But partly it is now seen to serve the ends of its own organizations. General Motors exists to serve the public. But General Motors also exists to serve itself as well or instead. Not many will find such a proposition radically in conflict with common sense. To quite a few it will seem trite. It is only remarkable in being at odds with the main thrust of economics as it is traditionally taught. (Galbraith 1973, pp. 19–20)[22]

The senior management of the large organization (what Galbraith memorably terms the "technostructure") has two fundamental purposes, one protective and one affirmative. The protective purpose is "to ensure a basic and uninterrupted level of earnings. Whatever serves this purpose – the stabilization of prices, the control of

[22] It is in Galbraith's work, as we shall see, that authentic Keynesian insights are integrated into an updated version of the institutionalist theoretical framework of Veblen and Commons. But of course, much progress was made in developing Keynes' insights before Galbraith, most notably by Michał Kalecki (who arrived at many conclusions similar to Keynes' at the same time or even earlier), Josef Steindl, Joan Robinson, and Sidney Weintraub. It is this Post-Keynesian tradition which Galbraith integrates with Institutionalism. For two excellent discussions of the place of Galbraith's work in the history of economic thought, see (Canterbery 1984; Dunn and Pressman 2005).

costs, the management of consumer response, the control of public purchases, the neutralization of adverse tendencies in prices, costs, or consumer behavior that cannot be controlled, the winning of government policies that stabilize demand or absorb undue risk – will be central to the efforts of technostructure and corporation" (Galbraith 1973, pp. 111–112). Fulfilling this purpose secures the continued existence of the technostructure and minimizes interference with its decisions. The notion that those who run large corporations need to maximize shareholder value in order to fulfill this protective purpose is groundless. "Given some basic level of earnings, stockholders are quiescent…proxy battles, when earnings are good, are virtually unknown" (Galbraith 1973, p. 110).

Nor is profit maximization the affirmative purpose of the technostructure. Instead, it is the growth of the firm. Growth serves "to reinforce the protective purpose of the technostructure…[and] also serves as nothing else the direct pecuniary interest of the technostructure" (Galbraith 1973, p. 116). This latter is so because "The contribution of any subordinate individual or group to earnings is merged with that of many others" and "[a]s a practical matter, no-one either inside or outside a company can tell whether profits are at a maximum. And there is no agreement as to the period over which profits are to be maximized" (Galbraith 1973, pp. 117, 123). On the other hand, "[i]n the case of growth…the contribution of the individual or a small group is often directly visible…growth often rewards directly those who are responsible for it" (Galbraith 1973, p. 117). Institutionalism, in virtue of recognizing the technostructure's real affirmative purpose, dovetails nicely with evolutionary economics, which models success within the marketplace in terms of growth rather than profits. What the evolutionary model does not capture on its own are the ways large trans-national corporations, as opposed to small firms, pursue growth. Following Galbraith, we may distinguish four of these: influence over prices, influence over consumer tastes, a type of innovation very different from that sought by small firms, and influence over public policy. Let us take these in turn.

4.1 Prices

Galbraith first notes that the role of prices within what he calls "the planning system"—the group of large trans-national corporations run by the technostructure which exerts oligopoly power—is very different than in the perfectly competitive market of neoclassical economics. In the neoclassical model, prices determine the allocation of resources. What determines the allocation of resources under the planning system is instead "the whole deployment of power – over prices, costs, consumers, suppliers, the government…The distribution of resources extensively reflects the power of the particular firm, and it is this power that allows us to speak of a planning system" (Galbraith 1973, pp. 127–128). The importance of prices to the planning system derives from the fact that these firms are engaged in a process of production which often requires complex and specialized capital equipment, and this is a source of large costs which are incurred before any product can be brought

to the market. In this high risk environment of large up-front investment, prices are important insofar as they are one factor—along with costs and consumer demand—which must be controlled to the greatest extent possible if the two purposes of the technostructure are to be served (Galbraith 1973, p. 129).

The large firms in the planning system, as members of an oligopoly, are not price-takers; and the level of their output has a significant effect on the price at which their products can be sold—Galbraith invites us to imagine the effect on the price of cars that a decision by GM to double its output would have (Galbraith 1973, p. 130). The ideal course of action for the management of the large firm, then, is to determine the price for its products which is most conducive to furthering its two purposes, and then set production based on how much can be sold at that price. Galbraith is very careful to note that the ability of a small set of large firms to approximate this course of action does not rest on the existence of explicit collusion:

> The power to set the price means that any other major firm in the industry…can, by fixing a lower price, force an alteration in the level first established. This may happen. But there is also a general recognition that such action, should it lead to further and retributive action by the firm originally establishing the price, could lead to general price-cutting. This would mean a general loss of control – a general sacrifice of the protective purposes of all the technostructures involved. The danger is recognized by all. In the planning system there is, accordingly, a convention that outlaws such behavior. It is almost perfectly enforced. No contract, no penalties, and usually no communication are involved. There is only an acute recognition of the disadvantage of such competitive and retributive action for all participants. (Galbraith 1973, pp. 130–131)

An arrangement of this sort, based on a tacit understanding by all parties of what is to be done and what the consequences are for failing to do it, is what economists refer to as an implicit contract. In consequence, the specific level at which prices are set will be determined by "the technostructure that is most committed to growth. Its price will be the lowest. Others must accept that price and therewith that goal" (Galbraith 1973, p. 132). The result of the exercise over prices by the planning system, then, is the opposite of what both neoclassical and Keynesian theory predict will be the result of a profit-maximizing oligopoly—lower prices and higher output, rather than higher prices and lower output, with extra-normal profits supported by the power to keep costs low through concentrated purchasing power and massive economies of scale—precisely because the affirmative purpose of the technostructure is maximum growth, not maximum profit. ("Lower prices" here means prices lower than smaller firms in a more competitive market could offer while staying in business; in a sense, oligopoly prices are higher than they should be, insofar as the members of an oligopoly do not compete with each other by trying to undercut each other on price, since each member knows that this would ultimately be harmful to itself. Instead, they compete through marketing and product differentiation.).[23] The

[23] Of course, smaller firms in more competitive environments use these techniques as well, though not to the same extent. Virtually all competition in modern capitalism is imperfect, and in particular, is between firms who do not provide goods or services judged to be perfectly substitutable by their customers—the demand faced by a firm is never perfectly elastic. Even in very competitive

real power of an oligopoly does not stem from its market-share, but from its ability to change the economic environment to suit itself. Since the public is inclined to approve of lower prices and greater availability of commodities, however, the existence of oligopoly and the exercise of oligopoly power does not often incite public outcry, despite the loss in efficiency and, as we will soon see, the negative long-term effects on innovation, economic stability, and overall economic growth (Galbraith 1973, p. 136).

4.2 Persuasion

The claim at the core of economic Institutionalism is that the preferences of individuals not only change (as even neoclassical economics will allow, so long as the changes are sudden and rare), and not only change gradually and continuously (as evolutionary economics will allow), but change endogenously. In particular, individual preferences are shaped and influenced by the institutions that structure the economic environment in which agents live (Hodgson 2000, p. 318). The influence of institutions on preferences is communicated through the control—which is often substantial even though it is by no means absolute—exerted over the structure of the choice-situations encountered by individuals: the range of options from which one can choose, the incentives or disincentives to make one choice or another, and the evidence and arguments offered (or withheld) concerning the values of the outcomes of those choices. Control over the first is achieved *via* the dominant market-share of the members of the planning system, which allows their productive choices to extensively influence what options are available in the marketplace; and, as we will soon see, *via* influence over public policy decisions. Control over the second is achieved through influence on prices. Control over the third is achieved through the exercise of persuasive power, primarily in the form of advertising.

To attribute significant persuasive power, exercised through advertising, to the firms in the planning system is not to make the absurd assertion that individuals confronted with corporate messages lose the ability to think for themselves, or that such messages are capable of shaping individual preferences completely. In fact, the impact of the persuasive efforts of any one firm in isolation, even a very large one, on individual perceptions and decisions may be quite limited. Rather, a significant degree of control over consumer preferences and demand is exercised, in a way that serves the purposes of the technostructures of all large firms, by the cumulative effect of all corporate persuasion taken together on shaping the default value system

markets, we no longer see the destructive price wars of mid-nineteenth century capitalism, a phenomenon which was integral to Marx's analysis and explanation of the trade cycle, particularly in *Capital: Volume 3*, Part 3, Ch. 15. See (Sardoni 2011, ch. 7). This stabilizing imperfection is largely the result of advances in business strategies which are characteristic of mature capitalism—just the sort of outcome which evolutionary economic theory would lead us to expect.

of a culture, which cannot help but exert a substantial influence on the judgments and choices of the members of that culture. As Galbraith explains:

> The advertising of the individual automobile company seeks to win consumers from other makes. But the advertising of all together contributes to the conviction that happiness is associated with automobile ownership…it encourages the general discarding of old vehicles and the purchase of new…More important still, the aggregate of all such persuasion affirms in the most powerful possible manner that happiness is the result of possession and use of goods and that pro tanto, happiness will be enhanced in proportion as more goods are produced and consumed. (Galbraith 1973, p. 156)

In virtue of its ability to recognize the existence, purpose, and importance of corporate persuasive power on individual preference-formation, Institutionalism helps us get a clearer view of the conditions required for autonomy-freedom. The development of autonomy is only possible within an energetically pluralistic and competitive society, in which no one institution or group of institutions holds an egregiously disproportionate amount of power to shape the range of available options, cultural values, and feasible life-paths. And so the possibility of an evolutionary market being populated by autonomous agents depends on the existence of an institutional structure which prevents the accumulation of too much economic and social power in too few hands. Within the liberal political tradition, the focus has historically been on the threat to liberty which is posed by a too powerful State. This is a vital concern, and I address the issue of the appropriate limits on State authority from the perspective of my theory in the final chapter. But the lesson of Institutionalism is that any organization, whether public or private, can pose this threat. While political liberalism requires a limited State, it also requires a State which is strong enough to prevent the emergence of corporate power on a scale that threatens the development of individual autonomy. I return to this point in the final chapter as well, in the context of refuting the claim that the liberal perfectionist State is necessarily objectionably paternalistic.

4.3 Innovation

In addition to pursuing growth through influence over prices and consumer preferences, there is no question that very large firms also do so through innovation. And some of the time, the sort of innovation they are motivated to pursue is the same as that pursued by smaller firms—innovation which responds to a perceived need or desire of consumers so far unfulfilled.[24] This is the sort of continuously adaptive innovation that the evolutionary model sees all successful firms as pursuing all the time, in order to remain successful, as well as all new firms eager for success. But

[24] There is a good deal of debate over whether it is large firms or small ones which are primarily responsible for beneficial innovation. This question, however, is probably a red herring, since in both cases the basic scientific research and technological development that forms the basis for subsequent private-sector innovation is primarily done by the State (Mazzucato 2013).

innovation within the planning system can, and often does, take on a very different form. Given their persuasive power and the role that many products play in conferring and communicating social status, the technostructure can pursue its affirmative purpose through innovation that produces novelty without increasing utility. "The popular view of invention," as Galbraith points out, "has long been strongly linear – there is a powerful presumption that a newly invented product is better…Given this view of invention, newness has sales value in itself. And this value persists, although perhaps with diminishing persuasiveness, even when there is no association between novelty and utility" (Galbraith 1973, p. 166–167). What supports this persistence is the importance of novelty in conspicuous consumption:

> [I]nvention in conjunction with advertising plays a vital role in stimulating the psychic obsolescence of goods and their replacement. This process…consists in creating a visually new product and then, through advertising, persuading the consumer that this is the only valid image of the product…The important thing is that the change succeed in making the earlier version visually eccentric and that its possession and use, in consequence, reflect discredit on the person so owning and using it. (Galbraith 1973, p. 167)

As novelty begins to outstrip increased utility as a source of greater sales, defect and malfunction become permanent fixtures of consumer experience (Galbraith 1973, p. 168). The logical end-point of these tendencies is of course planned functional obsolescence, which considerably eases the task of persuading consumers that the next generation of a product is superior to the last.

There are two further reasons, not considered by Galbraith, why the existence of oligopoly power is at odds with the emergence of beneficial innovation.[25] First,

[25] Despite his acute awareness of the threats posed by the planning system, Galbraith (like Schumpeter) believes that it does introduce a great deal of beneficial innovation and investment demand (which in turn contribute to greater economic growth, efficiency, and stability), and that its economic power can be contained without efforts by the State to limit the size of its members directly, by breaking up corporations which have grown dangerously large (Dunn and Pressman 2005, pp. 184–186). We are about to see, however, that as much as his theoretical system provides the most realistic framework for analyzing the workings of contemporary capitalism, these policy positions have been undermined over the last 40 years. Oligopoly power is now more highly concentrated than ever; governments have not shown themselves capable of limiting the power of large firms while allowing them to continue to grow; this ever-growing power has succeeded in undermining the countervailing force of the labor movement, which is crucial to Galbraith's model; investment demand generated by large firms has fallen steadily, creating a long period of economic stagnation and once again making the world's market economies vulnerable to regular asset bubbles; the largest corporations have achieved "systemic importance," compelling the State to prop them up in time of crisis with public funds; and this stagnation and instability decrease the rate of beneficial innovation and growth. For these reasons, I will base my outline of the policy program corresponding to the goals of Equal Liberty primarily on the modern German policy program of the Social Market Economy, which is far more pro-competition—not in the sense of seeking to create the conditions of neoclassical perfect competition, which we have seen is both a fool's errand and would not lead to the desirable outcomes claimed for it, but in the sense of actively seeking to curtail the emergence and persistence of oligopoly power. The Social Market Economy, to my mind, combines the most important insights of the authentic Keynesian and the institutionalist traditions in a better balance—though its authors arrived at their positions through an intellectual-historical path which had very little to do with Keynesianism or American Institutionalism.

oligopolies have an interest in protecting their business models by suppressing genuinely disruptive innovation; and second, global oligopoly power is the root cause of widespread financial crisis, which slows the development of all innovation. The interest of the members of an oligopoly in suppressing genuinely disruptive technological innovation is obvious: their profits and dominant market-share grow out of the success of the technology and methods they have developed and invested in in the past, and their success over the competition removes the pressure to innovate continuously. Suppressing disruptive innovation thus serves their protective purpose. Successful suppressions of disruptive innovation, however, are relatively rare. The only two important recent examples are General Motor's destruction of its own fully electric car technology in the 1990s, and the continued success of a number of large industry representatives to outlaw or otherwise discourage the cultivation and use of industrial hemp for the manufacture of paper, fabric, construction materials and biofuel.[26]

Far more important is the connection between oligopoly power on the one hand, and economic stagnation, financial crisis, and the resultant delay in the development of technological innovation on the other. The fundamental problem that arises along with highly concentrated global oligopoly power is the fact that for those who are realizing the greatest profits (the members of the oligopoly), opportunities for further growth through increased productive investment very quickly become scarce (because each member is already selling numerous goods across many industries and markets to virtually every household).[27] The possibility of new growth through increased productive investment is then tied to the occurrence of an economic shock—generally a new technological development which will lead to an even more efficient productive process for existing goods, new versions of existing goods, or an entirely new type of good and thus a new market. But major technological improvements take time to develop, and really influential ones are themselves rare occurrences. So some outlet is needed to absorb profits beyond what can be spent on new research and development without severely diminishing returns. And historically, a popular outlet has been speculation, brokered by financial firms, on short-term movements of asset prices. We have already discussed the role of what Keynes called "animal spirits" in the inflation of asset price bubbles. But the existence of highly concentrated global oligopoly power, along with recent developments in finance, exacerbates the psychological tendency to create asset bubbles. Modern finance is characterized by the construction and valuation of increasingly complex financial instruments (derivatives) whose value is derived from the value of some underlying real assets; and higher-order instruments (so-called "synthetic" derivatives), whose value is derived from the value of lower-order derivative instruments.

[26] On GM's electric car, see (Paine 2006). On the virtues of industrial hemp, see (Yonavjak 2013).

[27] Oligopolies also limit the opportunities for productive investment by ever other participant in the market, by concentrating so much capital into so few hands (Suarez-Villa 2014, pp. 229–230). In general, total productivity growth in major Western economies has been very slow since the mid-1970s, despite the emergence during this period of modern information technology—the great business innovation of the twentieth century (Lapavitsas 2013, ch. 7).

These developments in finance allow the total value of outstanding derivatives contracts to far exceed the total value of real production. The Gross World Product in 2007 was approximately 55 trillion USD.[28] The total value of outstanding derivatives contracts for the same year was 10 times that.[29] The existence of a highly concentrated global oligopoly drastically reduces the time interval between the development and widespread adoption of a (non-disruptive) technological innovation, and the point at which the new technology has been widely adopted and additional opportunities for investment are once again scarce. Global oligopoly thus produces longer periods in which large profits must be absorbed by activities other than new productive investment, and modern finance creates a virtually unbounded set of opportunities to channel those profits into asset-price speculation.

In time, profits from asset-price speculation begin to overshadow profits from real production—the market becomes "financialized."[30] There is then a temptation for dominant firms to direct profits away from new productive investment—or to take on ever-increasing levels of debt—to pursue additional speculation (Foster and McChesney 2012, ch. 2; Tabb 2012, ch. 2). This development results in a shift in the growth strategy which the planning system pursues with its profits: away from growth through increased production and innovation (since investment in these drops), and towards growth through market consolidation—the acquisition of competitors—which leads an oligarchic corporate structure to become even more highly concentrated. Oligopoly and financialization thus exist in a kind of feed-back loop (Suarez-Villa 2014, p. 87). In the U.S., financialization also created an irresistible incentive for the planning system to lobby for a vast expansion of access to debt and credit for the middle class. This boosted demand (and thus created some additional opportunities for corporate investment and growth in the short-term) through debt-financed increases in the standard of consumption (which generate greater profits for the financial institutions in the planning system) rather than wage-financed increases (which cut into the profits of non-financial corporations in the planning system) (Guttmann and Plihon 2010).[31] At the same time, large debt-equity ratios

[28] According to the World Bank.

[29] According to the Bank for International Settlements.

[30] In the 1950s, 15 % of profits at non-financial firms in the U.S. came from financial investments. That figure began to increase dramatically in the early 1980s, and by 2001, it stood at 50 % (Masouros 2013, p. 5).

[31] For an excellent historical narrative of the shift in the U.S. from a middle-class whose rising standard of consumption was supported by wage increases that tracked productivity increases, to one whose standard of consumption is supported by debt despite virtually uninterrupted annual increases in productivity, see (Hacker and Pierson 2010). It is no coincidence that the 1970s marked the beginning of the dizzying increase in income- and wealth-inequality in the U.S., which has resulted in the wealthiest 1 % and 0.1 % of Americans once again capturing a share of income and wealth on par with what those groups received in the Gilded Age of the 1920s. On this point, see (Piketty 2014, pp. 291–303, 314–321, and 347–350). This shift would not have been possible without the decline in the organized labor movement which occurred during the 1970s—a decline which was prompted by that decade's experience of "stagflation", which was in turn caused primarily by two factors: the supply shocks of the oil crises; and the excessively expansionary monetary policy of the late 1960s, which was motivated by an underestimation of the natural rate of

and the low interest rates of economic boom times allowed financial institutions to dramatically increase their own profits by engaging in speculation themselves with borrowed money. Joseph Stiglitz has recently argued that the expansion of the credit market has been the single biggest driver of wealth inequality in the U.S. over the last four decades, as the vast majority of the growth in wealth inequality during this period has been due to increases in the values of land and existing real estate, and the rental streams that flow to them, rather than in the value of the stock of moveable capital goods (Stiglitz 2015). This growth in the value of land assets has resulted from the loosening of rules limiting financial leverage, and has overwhelmingly benefitted the wealthiest individuals and corporations. Moreover, unlike increases in wealth due to increases in the value of the stock of other capital goods (which result from productive investment), this credit-fueled increase in the value of land assets does not contribute to economic growth and prosperity. Rather, it detracts from it, by concentrating the economic power of the wealthy holders of large amounts of land assets and encouraging them to engage in heightened levels of economic rent-seeking behavior (Stiglitz 2012, ch. 2).

It is the operation of "animal spirits" in this environment which led to the most recent financial crisis, whose severity was only curbed by action by public

unemployment, and early 1970s, which was motivated by a desire to stave off recession (Blinder 1982; DeLong 1997). The latter factor was a prime example of the neo-Keynesian policy of demand management in normal economic times which runs counter to the thought of Keynes himself. The same should *not* be said of Kennedy-Johnson era fiscal policy, and in particular of Johnson's War on Poverty, which did much to address economically harmful structural inequalities in American society. The 1960s saw the sharpest decline in poverty of any decade in American history, a decline which has been preserved to the present day, and Johnson's anti-poverty programs inspired a number of other successful programs implemented by subsequent administrations (Bailey and Danziger eds. 2013, ch. 4–8). At the same time, the Kennedy and Johnson administrations saw the greatest increases in real GDP *per capita* and real median income of any post-WWII American administration (and an increase in real median net worth which is a close second to that seen during the Clinton administration), as well as the greatest decrease in the annualized growth of the national debt as a percentage of GDP during the same period (Kimel and Kanell 2010). This is despite Johnson's great mistake – ensnaring the U.S. in the Vietnam War, which, in addition to its horrific human toll, cost the U.S. government $111 billion (approximately $700 billion in today's dollars). In addition, the rate of inflation actually remained under 5 % throughout Johnson's presidency. But with the retirement of ChairmanWilliam M. Martin from the Federal Reserve (the only effective brake on Johnson's desire for more expansionary monetary policy), and the Nixon administration's depreciation of the dollar after the collapse of Bretton-Woods, even looser monetary policy, and temporary price- and wage-controls in 1970 and 1971, still higher inflation was inevitable—the oil crisis of 1973–1974 notwithstanding. The first oil crisis exacerbated a situation which Nixon and his Fed Chair, Arthur F. Burns, had already worsened at precisely the moment when counter-cyclical action was needed, pushing inflation into double-digits. While subsequent Fed Chair Paul Volcker's policy of extremely high interest rates from '79 to'81 may have dampened inflation somewhat, this era of high inflation was ended by the steep decline in the price of oil which began in mid-1980, and which was quickly followed by the beginning of a significant decline in the rate of inflation. The subsequent drop in interest rates, coming a decade into the modern age of wage stagnation, provided the impetus for the current era of heavily debt-subsidized consumption among the lower- and middle-class. According to the St. Louis Federal Reserve, U.S. household debt as a percentage of disposable income rose from 68 % in 1980 to 128 % in 2007.

institutions to prop up aggregate demand in the face of massive destruction of household wealth,[32] and push the world's economies out of a liquidity trap in the face of a self-imposed freeze on lending by affected banks (Ferguson 2012, ch. 3–5). Sudden losses and a dramatic fall in aggregate demand strengthen the liquidity preference of established corporations, at least some of whom will already have been funneling profits away from new investment during the bubble. A credit freeze cuts off the operations of new small firms, including those in the process of developing new technology in an effort to break into a market dominated by a few very large organizations. Both of these events therefore have a large adverse effect on the very process of technological innovation which is required to create new opportunities for productive investment within an oligopoly system, and which is supposed to be the particular virtue of the evolutionary market which allows those oligopolies to arise. And they obviously lead to inefficiency, instability, and stagnation as well.

In fact, the move toward the financialization of the market has, for the past few decades, worked in concert with a misguided and ineffective attempt to align the interests of the technostructure with maximum profits rather than maximum growth, to produe economically disasterous results. In the U.S., the late 1970s saw the beginning of a shift in executive compensation from salary to stock options and stock awards. But rather than incentivizing executives to focus on profits, this shift, combined with changes made to SEC regulations under the Regan administration in 1982 which allowed corporations to make large open-market buy-backs of their own stock, led to an obsession among the members of the technostructure with short-term increases in share prices and quarterly earnings per share, at the expense of long-term profits and the interests of long-term shareholders (Lazonick 2014). (It was during the same period that a loosening of anti-trust rules paved the way for an explosion of corporate mergers and acquisitions.) The 449 companies listed on the S&P 500 which were publicly traded between 2003 and 2012 devoted 54 % of their earnings to stock buy-backs; another 37 % went to paying out dividends, leaving only 9 % (in addition to borrowed funds) for research, development, new capital investment and employee recruitment and incentivization (Lazonick 2014, p. 4). At the ten largest stock repurchasers from among this group, top executives received 68 % of their compensation in the form of stock options and stock awards (Lazonick 2014, p. 9). Open-market stock buy-backs are a means to creating short-term upward pressure on the value of a company's stock, for the sake of enriching senior executives whose compensation is overwhelmingly determined by share price and the extent to which it increases in the short-term. They are, in essence, a legal form of stock price manipulation (Lazonick 2014, p. 9). The dedication of profits to stock buy-backs rather than re-investment—profits which, increasingly, result from financial speculation or growth due to market consolidation, rather than growth in sales or marketshare due to innovation—further erodes the possibility of of increases in wages and employment levels, and thus of increase in the purchasing power of the middle-class, which is the *sine qua non* of long-term stable economic growth. The

[32] Median household net worth in the US fell 40 % between 2007 and 2010—setting it back to where it was in 1992. See *Federal Reserve Bulletin* June 2012).

technostructure's system of incentives could not be designed in a way that is more damaging to long-term broad-based prosperity.[33] And without broad-based prosperity, the goal of every individual enjoying an equal and extensive share of liberty is unattainable.

4.4 Public Policy and the Representational Limits of the Evolutionary Model

The last avenue to protection and growth which is exclusively available to the members of the planning system is the ability to use their vast wealth and economic power to bring public policy in line with their own interests.[34] Understanding this point is in fact the key to understanding the deep difference between neoliberalism, on the one hand, and classical liberalism and libertarianism, on the other. Neoliberalism has both an exoteric and an esoteric doctrine. Exoterically, it is an ideology founded on the same conservative conception of individual freedom, and faith in the virtues of free markets, as these other two views—indeed, it is difficult to distinguish from libertarianism, and both seem like little more than economically modern versions of classical liberalism. Esoterically, however, it takes an evolutionary-institutional point of view on the workings of a capitalist economy, recognizes the inherent potential for dominant firms to arise, and then sees influence over public policies and institutions as the spoils of that victory. As Philip Mirowski has passionately argued, neoliberalism, unlike classical liberalism and libertarianism, advocates an interventionist State (and cooperation in intervention across national governments). But the purpose of that intervention is to influence both markets and individual behavior in the interests of the largest firms—most notably, by enabling the expansion of credit-debt through low interest rates, and removing regulatory obstacles to leverage and speculation while providing implicit guarantees for those which are "too big to fail." The exoteric doctrine of freedom is used as a defense against calls for the State to act in ways that are counter to this purpose (Mirowski 2013).[35] Neoliberalism developed after WWII alongside the maturation

[33] See (Masouros 2013) for an outstanding study of the role of prioritizing the short-term, on the part of both executives and shareholders, in creating economic stagnation.

[34] For the reality and extent of this troubling phenomenon, see (Hayes 2013; Gilens and Page 2014).

[35] One could go a step further than Mirowski and argue that neoliberalism is really just a modern version of classical liberalism, because classical liberalism was characterized by the same duality. Its intellectual advocates may have been sincere in their belief that the exoteric doctrine of a night-watchman state which makes markets possible and then leaves them alone was the entirety of this school of thought; but in practice its conception of individual freedom was used to justify violent state intervention in the interests of business, most notably the forcible suppression of labor unionization. Perhaps all that is distinctive about neoliberalism is the specific way in which it has adapted the esoteric aspect of classical liberalism to modern times: abandoning open violence (at least within the developed world), and subduing a lower- and middle-class population which was thor-

of the planning system because it is the ideology of the planning system. And though it is often confused with the German theories of ordoliberalism and the Social Market Economy, this is a serious mistake. Ordoliberalism is founded on the idea that the appropriate role of the State is the preservation of the conditions of a reasonable level of competition, the prevention of great concentrations of economic power, and an absolute rejection of the translation of economic power into political power. The Social Market Economy adds to these commitments a recognition of the fact that there are many important aspects of a flourishing society which cannot be left to markets or in the hands of private enterprise, and thus that the role of the State includes the pursuit and maintenance of certain facets of the public good, and the counteracting of the social deficiencies of the market, through a variety of public institutions. We will take a closer look at the specific structure and policies of the Social Market Economy at the end of Chap. 11, where we will see that they are in close accord with the policy goals of Equal Liberty.

The manipulation of public policy by the planning system takes six main forms. The first is lobbying the government to erect regulatory barriers that limit the possibility of new firms entering and becoming competitive in the markets they dominate, or that force existing small firms to leave. The fact that US federal law prohibits the cultivation of industrial hemp without a DEA permit is an example of such a barrier, which serves the interests of a number of large industries. The second is lobbying the government to institute policies, or preserve privileges, that enable the extraction of economic rents. An individual or organization extracts an economic rent when it expends a portion of its resources for the purpose of shaping the economic or political environment in a way that will allow it to acquire a greater share of total existing wealth—i.e. a share greater than what it could expect in the context of a more competitive marketplace. This is in contrast to using resources to invest in productive activity which will create new wealth, a portion of which will go to the investor as profit. The paradigmatic example is acquiring publicly owned natural resources at well below market price, and reselling them on the open market at an inflated profit. The financial industry in the U.S. devoted considerable resources to rent-seeking in the years leading up to the financial crisis of 2007, by making financial markets less transparent and taking advantage of information asymmetries (Stiglitz 2012, pp. 37–42). The third is lobbying the government not to fund projects that serve the public good but conflict with their interests, or to fund projects that conflict with the public good but serve their interests. Lobbying by the automotive industry against government investment in more extensive public transportation systems is one example of this form (Galbraith 1973, p. 152). The fourth is lobbying against, and seeking trade agreements which prevent, the institution of regulations counter to their interests. The recent lawsuit filed by the Philip Morris corporation against the government of Australia, on the grounds that Australia's plain-packaging

---

oughly integrated into capitalism during the post-war boom through the cultivation of what Galbraith called "the convenient social virtue"—a view of the successful life which accords with the interests of the planning system—by replacing steadily rising wages with mounting consumer debt as the key to that success.

laws for tobacco products violates international trade treaties, is one chilling example. The global fight against internalizing the costs imposed by pollution from burning fossil-fuels, recently estimated at $5.3 trillion per year, is by far the most significant (Coady et al. 2015). The fifth is using influence over the legislative process to obtain subsidies including direct payments, tax incentives, and below-market loans and insurance plans—policy tools which should be used exclusively to support the small enterprises that constitute the market system. In the case of the U.S. fossil-fuel extraction industries, these amount to billions of dollars per year. The sixth is convincing the government to come to their aid in times of severe economic crisis, on the grounds of their systemic importance to the economic system. This enables very large firms to weather times of distress while smaller firms shutter their doors. The most recent examples in the US are of course the Troubled Asset Relief Program and the bailout of GM, Ford, and Chrysler.

We are finally in a position to understand how the emergence of stable, highly concentrated global oligopoly power from an initially competitive evolutionary market is possible. The evolutionary model fails to represent the various ways in which those firms which rise to dominance within their markets through competition have the potential to retain and fortify that position of dominance by altering the economic environment in which they operate to their own advantage—by influencing prices and consumer preferences; by generating new sales by exploiting competition for status, introducing novelty rather than beneficial innovation, and manufacturing obsolescence; and by influencing the content of the public rules and policies which constrain competition in the marketplace in ways that serve their purposes. Economic theory needs to go beyond evolutionary modeling, or even Keynesian agent-based modeling, and construct evolutionary-institutionalist agent-based models of the economy. It needs to explicitly represent the full range of public institutions—regulatory as well as monetary and fiscal—that shape the economic environment, establish the constraints within which market competition takes place, and participate in the market. And it needs to model the evolution of these public institutions themselves, as they interact with large firms seeking to influence the environment in which they operate. Finally, it needs to model the endogenous transformation of consumer preferences that result from interactions with large firms (a foundation for which is provided by my own work on preference change in combination with that of Dietrich and List).

Even more importantly for our immediate purposes, we can now appreciate the social limitations of a so-called free market—an evolutionary market likely to give rise to oligopoly power. This way of organizing the economy is not conducive to any of the social goals of efficiency, stability, beneficial innovation, or growth in productivity, employment, and household wealth. It is at odds with democracy, insofar as the technostructure succeeds in directing public policy decisions away from the wishes and interests of the majority of citizens. It is in conflict with the freedom of individuals to form their preferences autonomously, and limits the range of options from which individuals have to choose. Achieving even these basic socio-economic goals requires a system of strong public institutions capable of maintaining their independence from the private sector and pursuing policies that preserve reasonable

levels of competition and discourage the accumulation of oligopoly power. Such policies cannot fail to have the effect of limiting disparities in wealth—the primary way of preventing gross inequalities in power from emerging is by preventing gross inequalities in wealth. And we have already observed that even without assuming oligopoly power, policies which work to limit wealth inequality are required to tame the inefficiency and instability of the evolutionary market. So without positing equality of wealth itself as a social goal, we have arrived at the conclusion that curbing wealth inequality is of great social importance.

The story of social justice does not end here, with the goals of efficiency, stability, innovation, growth and democracy. It is only beginning. The next chapter is devoted to developing a theory of what full-fledged social justice requires: equality of liberty. Our conception of liberty is such that this goal is strongly incompatible with large inequalities of wealth and economic power, which severely limit the freedom of many to develop and exercise their autonomy while vastly expanding the freedom and power of few. But the immediate lesson is that even these antecedent social goals, goals typically associated with the operation of a free market, require strong, independent public institutions that prevent severe inequalities of wealth and power from emerging—the same sort of institutions which will be needed to achieve equality of liberty. The goal of equality of liberty and the five antecedent goals discussed in this chapter will thus prove to be consistent and mutually supportive. Before commencing this argument in the next chapter, I conclude my critical discussion of rival views of distributive justice by applying the lessons of this chapter to some of the most popular arguments in favor of a minimal State.

5 Against the Minimal State

The free market is the socio-economic arrangement which maximizes each individual's equal share of what we might call the conservative conception (or just as well the libertarian or classical liberal conception) of freedom: that is, the freedom to dispose of one's property, including one's labor, as one wishes. The minimal State is the politico-legal system whose sole function is the defense of individual rights to non-interference with the exercise of this freedom and thus of the free operation of the market. The argument against endorsing such a conception of freedom as our guiding moral and political value is the whole of Chap. 7 above. We already have in hand a very different conception of freedom, one which is rigorous, robust, philosophically well-motivated and, as I have argued at some length, ideally suited to playing that guiding role. Nonetheless, there are three important arguments defending the free market and the minimal State—and thus, defending the conservative conception of freedom that this arrangement makes preeminent—which must be countered, in order to clear the way for the chapters that follow. These are, first, that the free market protected by the minimal State produces the best socio-economic outcomes we can hope for; second, that the free market protected by the minimal State allocates to each individual the rewards he deserves for his activity;

and third, that any arrangement other than the free market protected by the minimal State entails the violation of individual moral rights.

5.1 The Axiological Defense

The argument against the axiological defense of the free market and the minimal State has actually already been given; all that remains is to point out this fact. The axiological defense claims that a market free of the influence of public institutions is the surest mechanism for generating efficiency, stability, and growth. These claims are based entirely on an approach to modeling the economy based on GET. As such, they are absolutely untenable. Strong, independent public institutions, acting both to shape and to participate in the market, are absolutely indispensable to achieving these goals. And from an evolutionary Keynesian or evolutionary-institutional perspective, it becomes clear that such institutions are indispensable even to the goal of innovation.

5.2 The Deontic Defense

5.2.1 Desert

The canonical version of the argument that the participants in a free market receive precisely what they deserve, whether in the form of wages or profits, is due to the late nineteenth/early twentieth century economist J.B. Clark. Clark's contention is that "what a social class gets is, under natural law, what it contributes to the general output of industry" (Clark 1891, p. 312). What he means by this is that the wage paid to the workers in a given industry is determined by the marginal productivity of labor in that industry, and the capitalist's rate of profit in that industry is determined by the marginal productivity of the capital goods purchased by the capitalist which he uses for production in that industry. The reward that each receives—the laborer in exchange for his hours worked, and the capitalist in return for the risk he assumed in making his capital investment—is thus a measure of his precise contribution to the process of production. This is the measure of desert, and a distribution in which each person receives what he deserves is a just one.

As in the case of the axiological defense, we already have all the tools we need to debunk this argument. The going wage only equals the marginal productivity of labor at a full-employment equilibrium in a neoclassical perfectly competitive economy—which is to say, never. And the relationship between the marginal productivity of capital and the rate of profit only holds at equilibrium in a neoclassical perfectly competitive single-commodity economy—which is to say, it does not

hold.[36] Clark's argument perishes along with the neoclassical framework in which he makes it. If wages in an actual economy do not depend on the marginal productivity of labor—and the empirical evidence is that they do not—what do they depend on?[37] The answer, which is as old as Adam Smith, is the relative bargaining power of workers and employers:

> What are the common wages of labour, depends everywhere upon the contract usually made between those two parties, whose interests are by no means the same. The workmen desire to get as much, the masters to give as little as possible. The former are disposed to combine in order to raise, the latter in order to lower the wages of labour. It is not, however, difficult to foresee which of the two parties must, upon all ordinary occasions, have the advantage in the dispute, and force the other into a compliance with their terms. (Smith 1776/2003, I.8)

In the present day, the power of U.S. corporations to influence legislation and public policy has resulted in rollbacks of workers' rights across the country, and a widespread unwillingness to monitor for and prosecute abuses (Lafer 2013).

But even if Clark's economic claims were correct, his argument would still fail. Whatever we may think about a theory of distributive justice based on some conception of desert, a desert-based theory of justice that equates the benefits one deserves with the measure of one's contribution to production is untenable. As the economist Joan Robinson explains:

> [T]he [orthodox economic] theory of distribution has nothing whatever to say, one way or the other, about the distribution of income. The theory purports to be concerned with the distribution of the product of industry between the factors of production. It says nothing about how the factors are distributed amongst people. The theory purports to explain the differences between skilled and unskilled wages, not how the chance to acquire skill is limited. It purports to explain rent per acre, not the size of estates; the rate of interest, not the possession of capital. (Robinson 1967, p. 75)

It therefore turns out—and this is a matter of no small philosophical interest—that Clark's argument must presuppose that individuals are morally entitled to their holdings of land and capital, and that limitations on opportunities do not result from

[36] In order to so much as *define* the marginal productivity of capital, it must be possible to quantify the total amount of capital. This is easily done if there is only one commodity—and thus one type of capital good—in existence. Given heterogenous capital goods, economists in the neoclassical tradition have wanted to use the total *value* of capital as a measure of the total quantity of capital. But the total value of capital depends on the value of a *prior* rate of profit. This makes it impossible to derive the rate of profit from the marginal productivity of capital, as the neoclassical school wants to do, and renders meaningless Clark's claim that profit is the reward deserved by the capitalist in return for saving, investing in capital goods, and putting them to productive use. This problem of capital-aggregation is at the heart of the so-called "Cambridge capital controversy," which occupied the economics departments of Harvard, MIT, and Cambridge in the 1960s. Many economists working in the neoclassical tradition to this day have never fully appreciated the problem. For the *locus classicus* of the critique of neoclassical production theory, see (Sraffa 1960). Andreu Mas-Colell has observed that Sraffa's critique parallels the other great aggregation problem that plagues neoclassical theory, the Sonnenschein-Mantel-Debreu theorem, in a number of interesting ways (Mas-Colell 1989).

[37] For the evidence, see (Frank 1994).

abuses of power. It must presuppose, in other words, that something like Robert Nozick's entitlement theory of distributive justice is correct, and that the world we live in is a just one by the lights of such a theory. It is to Nozick that we now turn.

5.2.2 Individual Rights

Serena Olsaretti has done an admirable job reconstructing Nozick's individual rights-based, or libertarian, defense of the free market and the minimal State (Olsaretti 2004, ch. 4–5). The official version of the argument can be stated as follows:

(1) A legitimate State is one which operates without violating any individual's moral rights.
(2) All individual moral rights are property rights: either rights of self-ownership or rights of external property justly acquired.
(3) The only involuntary justly enforceable obligations (or, equivalently, legitimate restrictions on freedom) are those correlative with other individuals' property rights. Call these "libertarian obligations."
(4) All other justly enforceable obligations are undertaken voluntarily, by voluntarily waiving some part or parts of one or more of one's property rights.
(5) To restrict an individual's freedom in the absence of either a libertarian or a voluntary obligation is to violate that individual's moral rights.
(6) In a free market with a minimal State, and only in this arrangement, freedom is restricted only on the basis of libertarian and voluntary obligations.
(7) Therefore, only in this arrangement does the State operate without violating anyone's moral property rights.
(8) Therefore, only the minimal State is legitimate.

Olsaretti's major contribution to debunking this argument is her astute observation that it relies on a highly implausible rights-based definition of the voluntary (Olsaretti 2004, p. 123). For the argument to go through, she points out, it must be the case that the wage-laborer who is faced with a choice between accepting a low-paying and physically-taxing job, or dying of starvation or exposure, accepts the job voluntarily, so long as (a) he has not been placed in that choice situation through actions of others that violate any of his rights, and (b) he is not coerced into choosing the job (which would also be a violation of his rights). But this is a hopeless definition of the voluntary. It would imply that one who has been justly imprisoned for a crime, and who is justly coerced into refraining from attempting to escape—justly because he has a duty not to, and no right to be free from this coercion—is in prison voluntarily (Olsaretti 2004, p. 125).

Voluntariness, then, does no work in the libertarian argument. All the libertarian means by voluntary action can be cashed out by specifying that the agent's rights—i.e. his property rights—are not violated in the determination of the choice-situations he faces, or in the making of his choices (Olsaretti 2004, p. 128). The libertarian argument can accordingly be reduced to the following:

(1) A legitimate State is one which operates without violating any individual's moral rights.
(2) All individual moral rights are property rights: either rights of self-ownership or rights of external property justly acquired.
(3) In a free market with a minimal State, and only in this arrangement, does the State operate without violating anyone's moral property rights.
(4) Therefore, only the minimal State is legitimate.

The first point to be made against this argument is that, given Nozick's own conception of moral rights, premise (3) is indefensible. Recall that Nozick can only establish the legitimacy (by his lights) of the minimal State by arguing that a universal prohibition by the ultra-minimal State on the use of private protective force does not violate anyone's rights, because private protective force poses such a significant risk of violating the rights of those it is used against. What Nozick neglects is the significance of the fact that in his state of nature, before the ultra-minimal State (or indeed any private protective alliance) is formed, this risk, though just as present, does not undermine each individual's right to use private force to protect his property against others. The reason for that is clear: there is no other alternative. The formation of the ultra-minimal State changes that fact. It becomes possible for rights to be enforced without posing such a significant risk of violating other rights in the process. Once this change has taken place, but only then, can it be said that a prohibition on private force does not violate the rights of the prohibited. It does not violate the rights of the prohibited because the right to use private force has ceased to exist, owing to the emergence of the ultra-minimal State. But forming the ultra-minimal State is a choice made by those who form it. Those who join the ultra-minimal State, then, by that very action, *take away* the right of those who do not join the ultra-minimal State, *without* the consent of the latter. Nozick completely misses this point when he describes forming the ultra-minimal State as within the rights of those who form it. That can only be true if those who do not join the ultra-minimal State lack an *immunity* which protects their right to use private protective force from being taken away by others without their consent. But Nozick has no way of defending this claim, and I cannot imagine it is even one he would want to make. His whole view is premised on the inviolability of moral rights. He cannot credibly turn around and claim that one's rights, though inviolable so long as they are intact, can be stripped through the unconsented-to actions of others.[38]

Nozick's reply would likely have been that the prohibition does not violate anyone's rights—not even immunity-rights—so long as it makes no one worse-off. And, so long as sufficient compensation is given to those who cannot afford to pay the State for protection, no one is made worse off. But this move is fatal to Nozick's defense of the minimal State, given the fact that the existence of a more-than-minimal State, taking policy actions which Nozick would want to count as

[38] This argument is analogous to Alan Gibbard's argument that no one in the state of nature can acquire unowned property without the consent of everyone else, since this act triggers a similar change in rights (Gibbard 1976).

rights-violations, can lead to Pareto-superior outcomes compared with a free market operating under a minimal State. Nozick's only alternative is to abandon the idea that one can "back into a State without really trying," and adopt a much more straightforward Lockean Contractarianism, according to which the minimal State is what all libertarian agents in the state of nature would choose to form, given their dominant interest in protecting the possession and integrity of their property.[39] He would then have to maintain that the only legitimate State—the only State whose operation violates no individual rights—is the one that libertarian agents would choose to form in the state of nature. It could be argued, moreover, that they have a moral duty to do so, since forming the minimal State renders the inherently risky use of private protective force both unnecessary and unjustifiable. We could then give an affirmative answer to Nozick's initial question: "If the state did not exist would it be necessary to invent it?" (Nozick 1974, p. 3) It would be necessary, not just in order to secure our interests, but in the moral sense of being what must be done in order to satisfy what we are bound by duty to do.

This revision does not save Nozick's theory, however. Rawls and Cohen have argued against the self-ownership thesis, the crucial element in premise (2) (Rawls 1971, §17; Cohen 1995, ch. 10). Cohen defines the concept of self-ownership thus:

> Each person…possesses over himself, as a matter of moral right, all those rights a slaveholder has over a complete chattel slave as a matter of legal right, and he is entitled, morally speaking, to dispose over himself in the way a slaveholder is entitled, legally speaking, to dispose over his slave. (Cohen 1986, p. 109)

The objection to the claim that each individual possesses this moral right of self-ownership is that one's capacity for productive activity, and thus for the accumulation of advantages through one's effort, is largely dependent on luck, both in terms of one's genetic inheritance and in terms of the environment into which one was born. Assuming that it is possible to formulate a principle of just acquisition of external property (and as many commenters on Nozick have pointed out, it is far from clear that he succeeds in doing so), the thesis of self-ownership affirms the justice of distributions which depend on morally arbitrary features of persons (a topic which was touched on in the previous chapter).[40]

The fact that the self-ownership thesis condones distributions that are significantly influenced by morally arbitrary features of persons is, to my mind, a powerful objection to it. But what no one, to my knowledge, has pointed out is the fact that the thesis of self-ownership, and the corresponding rejection of the claim that just distributions cannot depend on morally arbitrary features, are indefensible from a libertarian point of view. Nozick has a historical theory of what makes one's ownership of external property (i.e. property other than oneself) just. A holding of external property is just if, and only if, its original acquisition was just (and here Nozick attempts to formulate a principle of justly coming to own the previously unowned),

[39] This is the subtitle of Part I of *Anarchy, State, and Utopia*.

[40] For arguments against Nozick's principle of just original acquisition, see, for example, (Gibbard 1976; Cohen 1986).

or it was acquired through a just transfer (i.e. a transfer in which no one's moral property rights are violated) (Nozick 1974, ch. 7). However, the very concept of self-ownership, and of a moral right of self-ownership which is a property right, implies the existence of internal property as well as external; and part of one's internal property is one's share of the total of natural potential for productive activity in existence. By "natural potential for productive activity," I mean the greatest capability for productive activity which one could develop through one's effort, and the amount of effort that would be required to develop it, in ideal environmental conditions (nutrition, education, etc.) I take this to be a function of one's genetic endowment, and to be unequal across persons. Natural potential is one important factor in determining what actual level of capability for productive activity a person ends up developing, and thus in determining what share of external property one ends up acquiring, and passing on to one's descendants.

There is no discussion in Nozick's work of whether one justly holds one's share of internal property. Nozick simply assumes, as a brute moral fact, that every person has a moral property right to his actual share of internal property. But the distinction between external property and internal property is itself a morally arbitrary one. For the libertarian to be consistent, he would have to limit his affirmation of the moral right of self-ownership to those cases in which the process, whereby one acquired one's share of natural potential for productive activity, was a just one. But of course, this is nonsense. That process is the process of being born with a particular genetic endowment, produced and bequeathed by one's parents. The acquisition of one's genetic endowment, whether by transfer (*via* genetic inheritance from one's parents) or original (*via* mutation) is not the sort of thing that can be just or unjust (the idea of a just genetic inheritance would have to assume a just original acquisition at the origins of the family line, and again, this is nonsense). The lesson the libertarian wants to draw from this fact, combined with the fact that internal property is non-transferable, is that we have no choice but to simply posit that each person has a moral property right to his actual share. But this is a (rare) genuine commission of the naturalistic fallacy: there is no way to derive this moral conclusion from the mere fact that each individual finds himself with a given non-transferable share of natural potential.[41] Nozick's self-ownership thesis cannot be sustained, given his own historical theory of what makes property-holdings just. But the historical

[41] I refer here to Hume's naturalistic fallacy, whereby one deduces an 'ought' from an 'is', not to G.E. Moore's (alleged) naturalistic fallacy whereby one defines the property 'good' in terms of some natural property. It must be emphasized that one does not commit Hume's naturalistic fallacy whenever one asserts that the truth of some empirical proposition is relevant to the truth of some normative proposition. That is not fallacious at all. Hume's naturalistic fallacy is neither more nor less than to assert that: Things are in fact thus-and-so; therefore, it is morally right that things be thus-and-so. To commit the fallacy is to *moralize the status quo*. It is the oldest argument for cultural conservatism, and never should have survived the eighteenth century, given the availability of the non-fallacious Burkean argument from the wisdom of tradition and the risk of unintended consequences. The confusion about the meaning of Hume's fallacy likely stems from the fact that Moore's fallacy is a meta-ethical fallacy, and post-Moore interpreters of Hume have drifted into thinking of Hume's fallacy as also being meta-ethical. But it is not; it is a normative fallacy.

theory of justice in external property holdings relies in turn on the self-ownership thesis. If we do not have a moral right to our share of internal property on which our share of external property (partly) depends, then we do not necessarily have a moral right to our share of external property, even if it was acquired in accordance with Nozick's two principles of acquisition. And so too falls Nozick's defense of the legitimacy of the minimal State and the justice of the distribution achieved by a free market operating under a minimal State. A more robust State need not violate any individual rights; and a free market distribution may do so. The self-ownership thesis is the very foundation of Nozick's libertarian view. When it cracks, the entire edifice collapses.

Even if we reject this argument, and accept a moral right of self-ownership, Nozick's practical conclusions are still untenable. As Nozick would be the first to admit, what one may permissibly do with one's property is limited by the rights of others—just as the slave-owner may not, legally, have his slave murder his neighbor. But Nozick's claim that all individual moral rights are property rights is not based in any sound theory of moral rights; it is not based on any theory that provides a satisfactory explanation of why, and in virtue of what facts or features of the world, moral rights exist. Developing just such a theory is my goal in Chap. 14. And we shall subsequently see, in Chaps. 15 and 16, that the libertarian view of all moral rights as property rights is inadequate, and that the State whose authority can be morally justified is far from a minimal one. It is true that a legitimate State must operate without violating anyone's rights, and that if the State did not exist, we would have to invent it—in order to fulfill a number of the moral duties we owe to each other, as much as in any other sense. But the State we are referring to is not the night-watchman State. It is one which pursues the creation and maintenance of equality of liberty.

Finally, and perhaps most surprisingly, Nozick's libertarianism is practically inconsistent with the very possibility of a stable and efficient market, even given all the assumptions required by neoclassical economic theory. As we saw in our discussion of McCauley's work, among the many assumptions required to prove that a "free" market will arrive at a stable equilibrium is the assumption that its initial conditions lie on a stable asymptote. And by "initial conditions," we refer to the distribution of the natural endowment prior to the initiation of market activity. Nozick's libertarian society, then, could only include a stable and efficient market if, by some miracle, the distribution that resulted at the end of the process of original acquisition lay on a stable asymptote. Even if this is possible in principle, it is extremely improbable and its occurrence would be the purest accident.

References

Ackerman, F. 2002. Still dead after all these years: Interpreting the failure of general equilibrium theory". *Journal of Economic Methodology* 9(2): 119–139.

Admati, A., and M. Hellwig. 2012. *The Banker's new clothes: What's wrong with banking and what to do about it*. Princeton: Princeton University Press.

Arrow, K., and G. Debreu. 1954. Existence of an equilibrium for a competitive economy. *Econometrica* 22(3): 265–290.
Bailey, M.J., and S. Danziger (eds.). 2013. *Legacies of the war on poverty*. New York: Russell Sage.
Beinhocker, E. 2006. *The origin of wealth: Evolution, complexity, and the radical remaking of economics*. Cambridge, MA: Harvard Business School Press.
Berg, A., and J.D. Ostry. 2011. *Inequality and unsustainable growth: Two sides of the same coin?* IMF staff discussion note. Washington, DC: IMF.
Berg, A., J.D. Ostrey, and J. Zettelmeyer. 2012. What makes growth sustained? *Journal of Development Economics* 98(2): 149–166.
Bilancini, E., and F. Petri. 2008. A comment on Gintis' 'the dynamics of general equilibrium'. *Economic Bulletin* 2(3): 1–7.
Black, W.K. 2005. *The best way to rob a bank is to own one*. Austin: UT Austin Press.
Blinder, A.S. 1982. The anatomy of double-digit inflation in the 1970s. In *Inflation: Causes and effects*, ed. R. Hall, 261–282. Chicago: University of Chicago Press.
Canterbery, E.R. 1984. Galbraith, Sraffa, Kalecki, and supra-surplus capitalism. *Journal of Post Keynesian Economics* 7(1): 77–90.
Changes in U.S. Family Finances from 2007 to 2010: Evidence from the Survey of Consumer Finances. *Federal Reserve Bulletin* 98: 2, June 2012.
Chick, V. 1983. *Macroeconomics after Keynes: A reconsideration of the general theory*. Cambridge, MA: MIT Press.
Chipman, J.S. 1976. An episode in the early development of ordinal utility theory: Pareto's letters to Hermann Laurent, 1899–1902. *Revue européenne des sciences sociales* 14(37): 39–64.
Clark, J.B. 1891. Distribution as determined by a law of rent. *Quarterly Journal of Economics* 5(3): 289–318.
Coady, D., I. Parry, L. Sears, and B. Shang. 2015. *How large are global energy subsidies?* IMF Working Paper. Washington, D.C.: IMF.
Cohen, G.A. 1986. Self-ownership, world-ownership, and equality. In *Justice and equality here and now*, ed. F. Lucash. Ithaca: Cornell University Press.
Cohen, G.A. 1995. *Self-ownership, freedom, and equality*. Cambridge: Cambridge University Press.
Cohen, A.J., and G.C. Harcourt. 2005. Introduction: Capital controversy: Scarcity, production, equilibrium and time. In *Capital theory, volume I*, ed. C. Bliss, A.J. Cohen, and G.C. Harcourt, xxvii–lx. Northampton: Edward Elgar.
Crespi, B.J. 2004. Vicious circles: Positive feedback in major evolutionary and ecological transitions. *Trends in Ecology and Evolution* 19(12): 627–633.
Debreu, G. 1974. Excess-demand functions. *Journal of Mathematical Economics* 1: 15–21.
DeLong, J.B. 1997. America's peacetime inflation: The 1970s. In *Reducing inflation: Motivation and strategy*, ed. C. Romer and D. Romer, 247–276. Chicago: University of Chicago Press.
Dosi, G., G. Fagiolo, and A. Roventini. 2006. An evolutionary model of endogenous business cycles. *Computational Economics* 27(1): 3–34.
Dosi, G., G. Fagiolo, and A. Roventini. 2010. Schumpeter meeting Keynes: A policy-friendly model of endogenous growth and business cycles. *Journal of Economic Dynamics and Control* 34(9): 1748–1767.
Dosi, G., G. Fagiolo, M. Napoletano, and A. Roventini. 2013. Income distribution, credit and fiscal policies in an agent-based Keynesian model. *Journal of Economic Dynamics and Control* 37(8): 1598–1625.
Dunn, S.P., and S. Pressman. 2005. The economic contributions of John Kenneth Galbraith. *Review of Political Economy* 17(2): 161–209.
Edge, R.M., and R.S. Gurkaynak. 2011. How useful are estimated DGSE model forecasts? Washington, DC: Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board.

Ferguson, C.H. 2012. *Predator nation: Corporate criminals, political corruption and the hijacking of America*. New York: Crown Business.
Foster, J.B., and R.W. McChesney. 2012. *The endless crisis: How monopoly-finance capital produces stagnation and upheaval from the USA to China*. New York: Monthly Review Press.
Frank, R. 1994. Are workers paid their marginal products? *American Economic Review* 74: 549–571.
Frank, R. 2011. *The Darwin economy: Liberty, competition and the common good*. Princeton: Princeton University Press.
Friedman, M. 1953. The methodology of positive economics. In *Essays in positive economics*, 3–34. Chicago: University of Chicago Press.
Galbraith, J.K. 1973. *Economics and the public purpose*. New York: Pelican.
Gibbard, A. 1976. Natural property rights. *Noûs* 10: 77–86.
Gilens, M., and B.I. Page. 2014. Testing theories of American politics: Elites, interest groups, and average citizens. *Perspectives on Politics* 12(3): 564–581.
Gintis, H. 2007. The dynamics of general equilibrium. *Economic Journal* 117: 1280–1309.
Greenwald, B.C., and J.E. Stiglitz. 1986. Externalities in economies with imperfect information and incomplete markets. *The Quarterly Journal of Economics* 101(2): 229–264.
Guttmann, R., and D. Plihon. 2010. Consumer debt and financial fragility. *International Review of Applied Economics* 24(3): 264–282.
Hacker, J.S., and P. Pierson. 2010. *Winner-take-all-politics: How Washington made the rich richer—And turned its back on the middle class*. New York: Simon & Schuster.
Hayes, T.J. 2013. Responsiveness in an era of inequality: The case of the US Senate. *Political Research Quarterly* 66: 585–599.
Hirsch, F. 1976. *Social limits to growth*. Cambridge, MA: Harvard University Press.
Hodgson, G.M. 2000. What is the essence of institutional economics? *Journal of Economic Issues* 34(2): 317–329.
Hodgson, G.M. 2002. The evolution of institutions: An agenda for the future theoretical research. *Constitutional Political Economy* 13: 111–127.
Hoover, K.D. 1991. Mirowski's screed. *Methodus 3(1)*: 139–145.
Hotelling, H. 1935. Demand functions with limited budgets. *Econometrica* 3: 66–78.
Juglar, C. 1862. *Des Crises commerciales et leur retour periodique en France, en Angleterre, et aux Etats-Unis*. Paris: Guillaumin.
Kehoe, T.J. 1980. *Uniqueness of equilibrium in production economies*, MIT Economics Working Paper 271. Cambridge, MA: MIT.
Keynes, J.M. 1936/1964. *The general theory of employment, interest and money.* New York: Harcourt.
Kimel, M., and M. Kanell. 2010. *Presimetrics: What the facts tell us about how the presidents measure up on the issues we care about*. New York: Black Dog and Leventhal.
Kirman, A. 1992. Whom or what does the representative individual represent? *Journal of Economic Perspectives* 6: 117–136.
Kitchin, J. 1923. Cycles and trends in economic factors. *Review of Economics and Statistics* 5(1): 10–16.
Kondratiev, N. 1925. *The major economic cycles*. Moscow: Economica.
Koo, R. 2009. *The Holy Grail of macroeconomics: Lessons from Japan's great recession*. New York: Wiley.
Korotayev, A.V., and S.V. Tsirel. 2010. A spectral analysis of world GDP dynamics: Kondratieff waves, Kuznets swings, Juglar and Kitchin cycles in global economic development, and the 2008–2009 economic crisis. *Structure and Dynamics* 4(1): 1–55.
Krugman, P. 1998. *Japan's trap*. Available at http://web.mit.edu/krugman/www/japtrap.html.
Lafer, G. 2013. The legislative attack on American wages and labor standards, 2011–2012. Economic Policy Institute Briefing Paper #364.
Lapavitsas, C. 2013. *Profiting without producing*. New York: Verso.
Lazonick, W. 2014. Profits without prosperity. *Harvard Business Review, September 2014: 1–11*.

LeBaron, B., and L. Tesfatsion. 2008. Modeling macroeconomies as open-ended dynamic systems of interacting agents. *American Economic Review* 98(2): 246–250.
Mantel, R. 1974. On the characterization of aggregate excess-demand. *Journal of Economic Theory* 7: 348–353.
Mas-Colell, A. 1989. Capital theory paradoxes: Anything goes. In *Joan Robinson and modern economic theory*, ed. G.R. Feiwel. New York: New York University Press.
Masouros, P. 2013. *Corporate law and economic stagnation: How shareholder value and short-termism contribute to the decline of the western economies*. The Hague: Elven International.
Mazzucato, M. 2013. *The entrepreneurial state: Debunking public vs. Private sector myths*. London: Anthem Press.
McCauley, J.L. 2000. The futility of utility: How market dynamics marginalize Adam Smith. *Physica A* 285: 506–538.
Minsky, H.P. 1986/2008. *Stabilizing and unstable economy.* New York: McGraw Hill.
Mirowski, P. 1989. *More heat than light: Economics as social physics, physics as nature's economics*. Cambridge: Cambridge University Press.
Mirowski, P. 2013. *Never let a serious crisis go to waste: How neoliberalism survived the financial meltdown*. New York: Verso.
Myrdal, G. 1957. *Economic theory and underdeveloped regions*. London: Methuen.
Nozick, R. 1974. *Anarchy, state and Utopia*. Cambridge, MA: Harvard University Press.
OECD Directorate for Employment, Labour and Social Affairs. 2014. Focus on inequality and growth. Paris: OECD.
Olsaretti, S. 2004. *Liberty, desert and the market: A philosophical study*. Cambridge: Cambridge University Press.
Ostry, J.D., A. Berg, and C. Tsangarides. 2014. *Redistribution, inequality and growth*, IMF research report. Washington, DC: IMF.
Paine, C. 2006. *Who killed the electric car?*. Sony Pictures Documentary.
Piketty, T. 2014. *Capital in the 21st century*. Cambridge, MA: Belknap Press.
Polanyi, K. 1944. *The great transformation*. New York: Farrar & Rinehart.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Robinson, J. 1967. Marginal productivity. *Indian Economic Review (New Series)* 2(1): 75–84.
Saari, D. 1985. Iterative price mechanisms. *Econometrica* 53: 1117–1131.
Sardoni, C. 2011. *Unemployment, recession, and effective demand: The contributions of Marx, Keynes, and Kalecki*. Northampton: Edward Elgar.
Schumpeter, J.A. 1961. *The theory of economic development: An inquiry into profits, capital, credit, interest and the business cycle.* Trans. R. Opie. New York: Oxford University Press.
Scitovsky, T. 1976. *The joyless economy*. Oxford: Oxford University Press.
Slutsky, E. 1915. Sulla teoria del bilancio del consumatore. *Giornale degli Economisti e Rivista di Statistica* 3: 1–26.
Smith, A. 1776/2003. *The wealth of nations.* New York: Bantam.
Smith, J.M. 1982. *Evolution and the theory of games*. Cambridge: Cambridge University Press.
Sonnenschein, H. 1973. Do Walras' identity and continuity characterize the class of community excess-demand functions? *Journal of Economic Theory* 6: 345–354.
Sraffa, P. 1960. *Production of commodities by means of commodities*. Cambridge: Cambridge University Press.
Stiglitz, J.E. 2000. The contributions of the economics of information to twentieth century economics. *The Quarterly Journal of Economics* 115(4): 1441–1478.
Stiglitz, J. 2012. *The price of inequality: How today's divided society endangers our future*. New York: WW Norton.
Stiglitz, J. 2015. New theoretical perspectives on the distribution of income and wealth among individuals, part IV: Land and credit, *NBER Working Paper No. 21192*. Cambridge, MA: NBER.
Suarez-Villa, L. 2014. *Corporate power, oligopolies, and the crisis of the state*. Albany: SUNY Press.

Tabb, W.K. 2012. *The restructuring of capitalism in our time*. New York: Columbia University Press.

The Center on Budget and Policy Priorities. 2012. *The legacy of the great recession*. Available at http://www.cbpp.org/cms/index.cfm?fa=view&id=3252.

The Economic Policy Institute. *Issue Brief 367*. Available at http://www.epi.org/publication/ceo-pay-2012-extraordinarily-high/.

Title VII – Wall Street Transparency and Accountability, The Dodd-Frank Wall Street Reform and Consumer Protection Act. 2010. 124 Stat. 1376–2223.

Veblen, T. 1898. Why is economics not an evolutionary science? *The Quarterly Journal of Economics* 12(4): 373–397.

Vitali, S., J.B. Glattfelder, and S. Battiston. 2011. The network of global corporate control. *PLOS One* 6(10): e25995.

Yonavjak, L. 2013. "Industrial Hemp: A win-win for the economy and the environment" *Forbes.com.* May 29. Available at http://www.forbes.com/sites/ashoka/2013/05/29/industrial-hemp-a-win-win-for-the-economy-and-the-environment/.

Zarnowitz, V. 1996. *Business cycles: Theory, history, indicators, and forecasting*. Chicago: University of Chicago Press.

Chapter 11
The Theory of Equal Liberty

1 Introduction

The critical discussion completed in the last two chapters has directed us toward the conclusion that what distributive justice demands of us is that we work toward creating conditions in which every individual has equal freedom to develop and exercise his capacity of autonomy, and to act on the conclusions of that exercise by choosing which valuable capabilities to develop and exercise from among a broad range of options. This has served as a rough characterization of a theory of distributive justice. It is time to make it more precise. I call it the theory of Equal Liberty. It is a member of the equal opportunity family of views. It is an aspirational counterpart to Roemer's transitional account. And it reflects a basic commitment to the *Principle of Liberty*:

The primary goal of the state is to promote, preserve, and protect individual freedom and autonomy.

I begin this chapter by stating a principle which must constrain any candidate theory of distributive justice: the Principle of No Resource Waste. Any form of egalitarianism which is constrained by this principle is able to avoid some important problems. A commitment to the egalitarian ethos, moreover, makes satisfying the more demanding aspects of this principle less difficult. I then come to the heart of the chapter (and indeed of the whole book), in which I introduce the set of policy goals that characterize the theory of Equal Liberty. Each of the following subsections then provides a more detailed discussion of one of these policy goals, and responds to objections found in the philosophical literature against equal opportunity views in general. In each case, I argue that the objection does not undermine my account.

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_11

2 The Principle of No Resource Waste

I call the following the *Principle of No Resource Waste*:

(NRW) Do not waste resources that could be used to enhance the well-being, or the liberty to achieve well-being, of present or future members of society.

I take this principle to be a plausible deontic constraint on any adequate theory of distributive justice. That is to say, regardless of what the goal of one's theory of justice is, the pursuit of that goal must occur wholly within the bounds set by the principle.[1] The key to interpreting what the principle requires is of course interpreting the term "waste." There seem to me to be four important types of resource waste. First, there is resource destruction, wherein a portion of the resources at our disposal are simply obliterated. Second, there is what we might call "resource sinking," the use of resources in an attempt to achieve a goal which is known to be unattainable. Third, there is resource hoarding, wherein some power, most obviously the State, keeps some portion of public resources back from the populace and simply allows it to rot. This implies that the resources in question are not being invested for the future, are not regenerating (as might be the case for a public forest), and are not being held for the sake of future use. And fourth, there is insufficient conservation and policy investment.

I borrow Alan Jacobs' definition of a policy investment as a policy choice with the following two features: (1) Short-term aggregate resource extraction, whether in the form of taxation or of restriction on consumption; and (2) A mechanism of resource transfer toward the future, which may take the form of (a) accumulating resources for future consumption, (b) the creation of new goods with long-term value (such as new capital goods and public infrastructure), or (c) the production of new, slowly emerging consumption goods, of which conservation of natural resources is an example (Jacobs 2011, pp. 17–19). This explication of the term "waste" clearly needs to be filled in with more specific and concrete details. We need to know how we are to determine whether, for example, our current conservationist efforts are sufficient. In claiming that NRW plausibly constrains any distributive theory, I am asserting that each distributive theory must give some reasonable and precise answers to these questions. But the specific answers of each theory will likely be different, and influenced by the aims of the theory. I take responsibility only for providing answers to these questions from the perspective of my own specific egalitarian theory, which I will do in later sections of this chapter. Making a claim about the importance of investment and conservation for the future commits me to dealing with a suite of problems relating to the identity and well-being of future generations. I show how these issues can be dealt with from the perspective of my theory later in this chapter.

[1] One might object to the principle on the grounds that wasting certain resources might conceivably contribute to the well-being or liberty of some group of individuals, and that this contribution might outweigh the diminishing of the well-being or liberty of some other group which also results from the waste. But I do not think any plausible interpretation of well-being or liberty would sustain such a hypothetical counterargument, at least not from a multi-generational perspective.

Without getting into any further specifics, however, we can already see that an egalitarian theory is particularly well suited to satisfying the principle. The most demanding part of the principle is surely the last, which requires conservation and investment in the present for the sake of the future. The goal of Jacobs' work is to uncover the necessary conditions for governments to willingly and successfully make policy investments. His research has identified three such conditions (Jacobs 2011, p. 23). The first is electoral safety. Policy investment does not occur when lawmakers fear electoral punishment for the imposition of the required short-term costs. The second is confidence in social benefit. Lawmakers only make policy investments when they are quite confident that those investments will have positive returns in the long-run. The third is what Jacobs calls "institutional capacity," and this refers to an absence of resistance on the part of large organizations composed of those who would bear the brunt of the short-term costs. These conditions are least likely to be met in a society where there is great inequality of economic power.

In the U.S., the power of individuals and the organizations that represent them to influence the policy outcomes generated by the legislative process is highly correlated with the relative wealth of those individuals (Hayes 2013; Gilens and Page 2014). And the wealthier they are as compared to their fellow citizens, the more disproportionately the short-term costs of policy investment can be expected to fall on them. For the legislature that overcomes such resistance, electoral safety is sure to be a major concern, and more so in a more unequal society. The penalty is likely to be a withdrawal of financial support—more difficult to offset when wealth is concentrated in fewer hands—and a generously bank-rolled challenger in the next election. Great inequality even threatens the second of Jacobs' conditions. As Jacobs notes, "long-term policy consequences typically depend on highly complex causal processes" (Jacobs 2011, p. 23). Accurately forecasting long-term policy consequences often requires the use of sophisticated scientific research. The disproportionate influence of a small number of extremely wealthy individuals over the legislative process, however, can lead lawmakers to exclude the research needed for accurate policy forecasting when that research points in a direction unfavorable to powerful interests. This is precisely what happened in North Carolina in 2012, when the state legislature passed a law prohibiting the use of existing data on the effect of climate change on coastal areas for the purpose of regulating coastal development until 2016 (North Carolina House Bill 819 Section 2 2012). There is good reason to believe, then, that the very possibility of policy investment, and thus of satisfying NRW, requires that electoral and legislative processes be insulated from concentrated economic power. We will return to this point at the end of the chapter.

From a single-generational perspective, NRW may seem to be at odds with the Pareto Principle, which states, roughly, that we should always prefer a Pareto-efficient allocation of resources to a non-Pareto-efficient one. An allocation of resources is Pareto-efficient when there is no agent who can be made better-off through a transfer of goods that does not also make some other agent worse-off. NRW requires that some resources be held back, and this implies that there is at least the possibility that some members of society could be made better off without

making any other (currently existing) members worse-off. But NRW is meant to be a principle of multi-generational distribution; and indeed, I think the Pareto Principle also ought to be so interpreted. NRW is meant to complement the Pareto Principle. It is possible to satisfy NRW without satisfying the multi-generational Pareto Principle, and vice versa. So NRW may be seen as a guide to selecting a particular multi-generational Pareto optimum. It gives us ground to exclude some multi-generational Pareto optima as inconsistent with distributive justice. Which ones it guides us to will vary with our interpretations of insufficient policy investment. As I said above, I address this issue later in this chapter.

NRW is significant for two other reasons. The first is that many versions of egalitarianism (though not my own) are inconsistent with the Pareto Principle, and NRW provides them with a principle to accept in its place that captures some of what makes it attractive.[2] The second is that, whether one's theory is consistent with the Pareto Principle or not, not everyone finds it appealing. For these too, NRW offers a replacement. Amartya Sen famously argued that Pareto efficiency is inconsistent with a commitment to what he calls Minimal Liberalism: the existence for each agent of a purely personal sphere in which the agent is able to satisfy his preferences without the threat of interference from any other agent. Sen refers to this result as "the impossibility of the Paretian liberal," and has ably defended it against a number of attempts to diffuse it (Sen 1996). I discuss Sen's principle of Minimal Liberalism in the introduction to the next part of the book. My own view regarding the liberal paradox is that we should interpret the Pareto Principle not as concerned with actual preference satisfaction, but as concerned with well-being—understood according to the theory of well-being already developed. The satisfaction of anti-social preferences—the sort of preferences agents must have in order to generate the paradox—does not count as a contribution to well-being according to this view, and so the Pareto Principle remains consistent with a commitment to Minimal Liberalism. This of course is not an original solution to the paradox. The idea that we should challenge Sen's domain assumption, which is what allows for anti-social preferences to count in social decisions, is practically as old as the paradox itself. But what is required of anyone who takes this route is that he provide a robust theory of well-being which restricts Sen's domain of admissible preferences in a principled and plausible way. I believe that my theory of well-being does this. Nonetheless, for those who would rather reject the Pareto Principle, they may find in NRW a replacement principle which is consistent with Minimal Liberalism even on Sen's domain.

By accepting NRW, even the strict egalitarian is able to avoid a number of objections to his view. First, there is the "leveling-down" objection. We were asked to consider a society with two groups of equal size, A and B, and the following two possible distributions:

(1) Each member of group A—1
Each member of group B—1

[2] For the incompatibility of many forms of egalitarianism with the Pareto Principle, see (Tuggoden and Vallentyne 2010).

(2) Each member of group A—9
Each member of group B—8

At this point, we need to ask a question which is insufficiently emphasized in the literature on prioritarianism: *What would we have to do in order to achieve distribution (1), given that distribution (2) is attainable?* I submit that there are two answers to this question. Either (a) a large portion of the resources at our disposal would have to be destroyed, so that what was left was just enough to effect distribution (1); or (b) that same portion of our resources would have to be hoarded by the State. Both of these are prohibited cases of resource waste. The egalitarian whose pursuit of equality is constrained by NRW must prefer to bring about distribution (2). Of course, we can interpret the leveling-down objection in another way. The point of the objection might be that, presented with two possible worlds represented by distributions (1) and (2), and *assuming that in the world represented by (1), distribution (2) neither is nor was attainable*, the strict egalitarian must *judge* world (1) *better* than world (2). This is a much weaker objection; it is unclear whether it has any practical import. But I am content to let it stand against the strict egalitarian.

Two other anti-egalitarian arguments of Arneson's, which do not only affect the strict egalitarian, can also be refuted. The first involves the two island populations. Consider a group of badly-off individuals who are in possession of some crop seeds which cannot grow in the soil where these individuals live. The seeds, however, would sprout in the soil where a group of slightly better-off individuals live. Assume it is possible to transport the seeds but not possible to relocate the badly-off individuals, and that the slightly better-off have nothing they could afford to part with in exchange. If it really is the case that the seeds are of *no* use to the badly-off, then giving the seeds to them, or allowing them to keep the seeds, counts as resource waste; it is tantamount to hoarding. The egalitarian who accepts NRW then has no problem with this case. The second involves two groups, one of which is just above subsistence, and the other of which is below, and in danger of beginning to die off. Assume that we cannot transfer enough from the better-off to keep them above subsistence while also raising the worse-off above subsistence. An equal distribution would then condemn all to death. However, if we do not possess enough resources to bring an individual above subsistence, and have no reason to believe that we will in the near future (near enough so that he will still be alive), then the transfer of resources to him also counts as resource waste—in the form of resource sinking—though we should perhaps strive for gentler language to describe such a tragic scenario.

3 The Goals of Equal Liberty

The specific policy goals of Equal Liberty reflect the elements of a life of liberty, as presented in the account of individual liberty given at the end of Chap. 7. They are as follows:

1. Equality of basic functioning
2. Equality of freedom and encouragement to develop and exercise the capacity of autonomy.
3. Equality of freedom for autonomous capability choice and development.
4. Equality of capability subject to effort and capability choice.
5. Equality of freedom for capability exercise
6. Maximin of achieved functioning subject to autonomous effort and voluntarily accepted risk.

These goals are ranked in order of priority; but the ranking cannot be taken to be a strict lexicographic one. The reason is that the pursuit of each is restricted by the NRW Principle. In the event that, owing to the circumstances prevailing at a given time, a higher priority goal cannot be pursued at that time without violating the NRW Principle, the next highest priority goal whose pursuit is consistent with the principle at that time is to be pursued. The NRW Principle thus serves as both a deontic constraint on the operation of the State in general and as a principle of trade-offs between goals within the Equal Liberty framework.

The successful pursuit of the policy goals of Equal Liberty is sure to require a robust economy capable of significant wealth creation. The basic economic goals of efficiency, stability, growth and innovation are therefore prior to the policy goals of Equal Liberty: Equal Liberty is a goal that can only be sought by a society with a well-functioning economy. Note that this is a purely instrumental priority: the morally more important goals of Equal Liberty cannot be effectively pursued and achieved, over the long term, in the absence of a healthy economy. We need not rank the basic economic goals. As we saw in the last chapter, they are largely complementary and mutually reinforcing. There is a great deal of overlap between the sets of policy actions that promote each of them, and achieving any one contributes to achieving the others. We can therefore rank basic economic policy programs into indifference classes, where any program that leads to inferior performance in the pursuit of all these goals ranks below any program that leads to superior performance in all categories. The complementary and mutually reinforcing nature of the goals themselves will prevent wide divergences in performance in any one category among the policy programs within a given indifference class. There might seem to be a danger of incompatibility, or at least of tension, between these two sets of goals. The economist Abba Lerner, for example, posits that in a dynamic production economy "the principle of equality would have to compromise with the principle of providing such incentives as would increase the total of income [or resources generally] available to be divided" (Lerner 1944, p. 36). However, if the "principle of equality" in question is a commitment to Equal Liberty, as opposed to a principle of equality of welfare, then we need not view this relationship as one of compromise. We shall see, building on the results of the last chapter, that the sorts of policy actions that actually promote the basic economic goals all promote the goals of Equal Liberty as well; and the further policy action required to pursue the further goals of Equal Liberty are, for the most part at least, consistent with pursuit of the basic economic goals.

Finally, the pursuit of those basic economic goals, no less than the pursuit of the goals of Equal Liberty, is constrained by the NRW Principle. Obviously, it is the fourth part of the principle that is important here. What the NRW Principle actually does in this context is force a particular interpretation of success in the pursuit of efficiency, stability, innovation and growth. Success must be evaluated from a long-term perspective. We seek long-term, ecologically sustainable efficiency, stability and growth, and the innovation of ecologically sustainable technologies. One set of policies surveyed in the last chapter as essential to these goals in general, as it turns out, is particularly well suited to their sustainable achievement. The sort of growth and innovation achieved in an environment of highly concentrated oligopoly power is, as Galbraith observed, the sort that "adds to pollution of air and water and to other environment disharmony… It is the pursuit by the technostructure of its own goals, exercising its own power to do so, and not technological innovation per se, that is at the heart of the environmental problem" (Galbraith 1973, p. 166). A recent report prepared for the United Nations estimates that the world's largest 3000 firms are responsible for 2.2 trillion USD worth of damage to the environment, and that internalizing these costs would reduce total profits by 33 % (UNEP Finance Initiative 2010). An absence of oligopoly power, as we have seen, is one of the essential features of a stable, efficient, robust and innovative economy. One of the most effective weapons against the emergence of oligopoly power is the internalization, through rigorously enforced regulative policy, of the costs of natural resource extraction and of cleaning up and preventing pollution.[3] And this, combined with incentives for the development and use of sustainable technologies, is the key to long-term environmental conservation. So there is a strong complementarity and mutual reinforcement here as well, between adherence to the NRW Principle and the pursuit of the basic economic goals which that principle constrains.

We now proceed to examine each of the policy goals of Equal Liberty in greater detail. In what follows I will often use the expressions "opportunity for development" and "opportunity for functioning/well-being" rather than "freedom to develop" and "freedom to achieve functioning/well-being" when the former phrasing is more natural. In either case, the conception of freedom developed in Chap. 7 applies.

3.1 Equality of Basic Functioning

The first goal of Equal Liberty is the provision of resources required for basic functioning to all members of society. I assume that there exists, at least for a given society at a given point in history, a set of functionings which could be called "basic," in the sense that they are the successful exercise of "the (nearly) universal capacities, i.e. those necessary for all or almost all valuable pursuits" (Raz 1994, p. 17). The task of enumerating the basic functionings for the members of a

[3] For policy and implementation strategies, see (UNEP 2008).

particular society at a particular time is not one for philosophy, but for the combined resources of the biological and social sciences. Nonetheless, this set is sure to include exercise of "the basic physical and mental abilities of controlled movement and, where disability deprives one of them, appropriate substitutes," and of the mental abilities required to "form, pursue and judge goals and relationships"—that is to say, the basic rational capacities which are prerequisite for the development of autonomy (Raz 1994, p. 17). It includes the exercise of some basic level of self-control. It also includes being adequately nourished, being protected from the elements, and, in many societies, being literate and having basic mathematical skills.[4] Finally, for those who have achieved its other components, basic functioning includes attaining a level of dignity. One has dignity in the relevant sense when one perceives oneself, and is perceived by others, as a person who has the potential to make autonomous life-choices and to participate in civil society and the political community. Dignity, therefore, is achieved through the establishment and recognition of a maximally extensive scheme of equal civil and political rights, which enshrine the equal social standing of members of different genders, ethnicities, religions, economic classes, and other groups which help to constitute one's personal identity (Kenny and Kenny 2006, ch. 5–6).

What counts as the achievement of basic functioning will be position dependent for a given individual, as well as relative (to an extent) to one's society and historical period. For the severely developmentally disabled, for instance, it may be impossible for the realization of the mental abilities required to form, pursue and judge goals and relationships to extend past the point where they would in a non-disabled child of 4 or 5 years. These individuals are to count as having achieved basic functioning when their mental abilities are developed to the point that matches the age of mental maturity that it is possible for them to reach, as determined by appropriate diagnostics. But we will have failed them if we have not equipped them with the resources, aid and environment required for them to achieve the level of development that is possible for them. In some cases, individuals are so severely disabled that they cannot hope to achieve anything like the mental functioning that Raz describes. It seems that we are forced to judge that a truly human way of life is beyond the reach of these individuals. Our responsibility to them is to provide them with the resources, aid and environment they require in order to lead lives of as much activity and as little pain as is possible for them.[5]

This first goal may seem open to the following objection:

[4] Rawls eventually recognizes the importance of a principle of justice which guarantees the satisfaction for all individuals of what he calls "basic needs," like nourishment and shelter, and gives it lexical priority even over the Principle of Liberty (though he could alternatively have insisted that the Principle of Liberty must itself be interpreted as mandating the satisfaction of those needs which make the exercise of political and civil liberties possible) (Rawls 1987, p. 166).

[5] The fact that this is a commitment of my view shows it to be a form of perfectionist liberalism capable of meeting David Charles' second challenge. See the discussion of Charles in Chap. 9 *supra*.

The Ascetic Life: A liberal society should respect the choices of those who willingly lead lives of spiritual asceticism, even though such forms of life may include or lead to undernourishment and other physical impairments.

This objection, however, does not undermine a commitment to equal basic functioning. The key point is that what must be respected, as the objection acknowledges, is the *choice* to lead an ascetic life. But for the agent to *choose* asceticism, a choice which should be respected (and for which he should be held responsible), the agent must possess and exercise the mental abilities necessary for freely and deliberately making that choice. And developing those mental abilities requires a basic level of physical health and wellness (Watanabe et al. 2005). Basic functioning is required for all of the life choices that liberal society recognizes as worthy of respect. We simply must assume, therefore, that an individual who has never achieved the basic level of functioning has not chosen a life of saintly denial, but has rather been unjustly deprived of the first requirement of a valuable life. And our first goal should be a distribution of resources, services, hands-on aid and education that results in the achievement of basic functioning for every individual.

But why should *equal*—rather than, say, *maximin*—basic functioning be our goal. The simple answer is that I do not see basic functioning as something that can be maximized. There is a (position-dependent) level at which an individual can be said to be achieving basic functioning. Achievements beyond this level are not basic. There are, in other words, degrees of failure to achieve basic functioning; some individuals are much worse off than others, even though none in the group may be said to achieve basic functioning. But there are no degrees of achievement. Of course, we need some way to rank distributions of individual levels of basic functioning, so that we can say, of any two such feasible distributions, which one brings us closer to our goal of full achievement of basic functioning by all individuals. And the social choice rule we use to establish such a ranking of distributions must reflect our commitment to egalitarianism. I will address this issue in Sect. 3.6.2 below, where I treat the topic of egalitarian social choice rules.

3.2 Equal Opportunity and Encouragement for Autonomy Development

William Galston has argued that liberal societies are concerned with three aspects of the lives of their citizens: need, desert, and choice (Galston 1991, p. 184). What has emerged from our discussion so far is that these three aspects are interrelated in a number of ways. The concern for basic functioning corresponds to need. The provision of basic functioning is a necessary prerequisite for the ability to make choices for which one can be held responsible, including choices about what risks to accept, and the level of effort to exert in pursuit of one's goals. The latter is the main determinant of the level of well-being one deserves to achieve in one's life. But even though one may have achieved basic functioning, one may still be incapable of

making choices for which one can be held fully responsible. The ability to make such choices is an ability one must cultivate, and its cultivation requires that basic functioning already be achieved.

The theory of Equal Liberty sees the distributive goal of the State as that of creating the conditions in which individuals are capable of making choices for which they are responsible, and in which the level of well-being achieved by an individual will be determined by the level of effort he freely and autonomous puts into his own life, subject to the risks he freely and knowingly accepts. After basic functioning, then, the theory is concerned with the distribution of the freedom to develop the capacity of autonomy—the capacity to make choices about what goals to pursue, and with what level of dedication to pursue them—and with the distribution of encouragement for this development. Agents must possess this capacity in order to exercise the freedoms which a liberal society affords them in a way for which they are responsible.

The freedom to develop one's capacity of autonomy for those who have achieved basic functioning is only available within a society that recognizes the importance of, and endorses, competitive value pluralism. As Stanley Benn has argued:

> It appears, then, that autonomy is an ideal available only within a plural tradition, for it requires that two conditions be satisfied. In the first place, it requires that the subject's beliefs be coherent and consistent; secondly, their coherence must be the outcome of a continuing process of critical adjustment within a system of beliefs in which it is possible to appraise one sector by canons drawn from another. A monolithic system, in which, for instance…ways of acting have been routinized by a kind of natural selection process for all major eventualities…would simply lack the incoherences which leave space for autonomous development. (Benn 1988, p. 182)

Benn takes autonomy to be an ideal, a perfection attained, and so requires that the autonomous agent have a system of belief which is perfectly consistent and coherent. I have developed a theory of autonomy that takes the autonomous agent to be one who is filling in a conception of the good life in a certain way, strengthening its coherence and consistency as he goes—the sort of agent that Benn would refer to by the unlovely term "autarchic." Benn's second condition, however, is a nice statement of the environment that is required for the development of autonomy. It fits nicely both with my view of autonomy and my theory of distributive justice. We develop autonomous preferences as a result of conducting what Mill called "experiments in living." We can only conduct these experiments if we have access to a suitable laboratory. The competitively pluralistic society is that laboratory.

In the first place, it is the fact that society endorses competitive value pluralism that offers agents the freedom they require to develop the capacity of autonomy. Autonomous reasoning about ends, as I have modeled it, is not a unique or exotic sort of reasoning. It is structurally quite similar to instrumental reasoning, and requires the same sort of responsiveness to evidence that is required in developing any informed beliefs about any aspect of the world. Those who are capable of reasoning instrumentally and of responding to evidence that bears on their beliefs are also capable of reasoning about ends. And the mental fitness required for these abilities is a good candidate for inclusion in the set of basic functionings (at least in

cases where severe disability is not present). But also required for autonomous preference formation is access to data. Excellent reasoning about ends requires information about potential ends, and truly autonomous preference formation requires access to a good deal of information about a wide variety of potential ends. This information is gathered both through one's own experiments in living, and from the testimony of others. The more freedom an agent has to experience at least a taste of a variety of valuable pursuits that could be incorporated into his own conception of the good life, and the more freedom others have to do the same, and then share their evaluations of their experiences, the greater the agent's opportunity to develop his own capacity of autonomy.[6]

The development of autonomy requires the presence in society not only of pluralism, but of competitive pluralism. And this is for the reason Benn identifies in his second consideration. Suppose one were to live in a society that permitted its members to pursue a variety of values, but that the permitted values were all complementary, in the sense that they could all be pursued and realized within a single life. Such a society would eliminate the need for its members to make trade-offs and compromises—to confront the options that are available to them, and work to determine what is really important to them and what they ought to forgo. This process, however, of trade-offs and compromises, of dealing with conflict between values and determining how to structure one's life based on one's conclusions about what is really important to one—this is precisely what the process of autonomous preference formation amounts to. It is a process of conflict resolution—conflict between the claims of competing specifications, conflict between old ends and new ones, conflict between established preferences and new evidence that threatens to upset them. It is thus by building a society that recognizes, respects and endorses competitive value pluralism that we create the conditions in which agents are free to develop the capacity of autonomy.

The second policy goal of Equal Liberty, however, is not merely equality of opportunity to develop one's autonomy. I have also included equal encouragement of that development. Suppose we succeed in organizing society in such a way that every individual achieves basic functioning (including developing the mental abilities required for autonomous preference formation), and every individual enjoys equal freedom to develop his autonomy. There is equal access to the data required for forming autonomous preferences, as each individual has a broad range of opportunities to expose himself to valuable pursuits that he could incorporate into his plan of life, as well as access to the views, judgments and evaluations of others regarding their own experiments in living. Despite having the necessary abilities and access to the necessary resources, some individuals may still fail to develop into autonomous agents. This might happen even though there are no external obstacles to their development—physical, social, cultural, economic, or otherwise. I argued above that the liberal egalitarian should be concerned with the freedom to achieve

[6] Autonomy is therefore a capacity whose healthy development depends in no small part on our relationships with others. For a discussion of autonomy that focuses on this feature, see (Nedelsky 2011).

well-being, rather than with the achievement itself, because the liberal State should aim to create conditions in which individuals are responsible for the effort they put into their own lives, and thus deserve the well-being they achieve. In the case of autonomy development, however, freedom is not enough. On the Scanlonian view of responsibility that I endorse, autonomy, at least as I have defined it, is precisely what is required for responsibility. So actively encouraging the development of individual autonomy is part of creating the conditions that the liberal State strives for, and even in the absence of any external obstacles it is imperative that autonomy development be actively incentivized. The State will have achieved this goal precisely when all individuals achieve the highest level of autonomy which is feasible for them—as measured according to one of the approaches mentioned at the end of Chap. 7. There can be no argument against such encouragement on the grounds that any individual is responsible for not cultivating his autonomy. The very idea that someone is responsible for a failure presupposes that that failure results, at least in part, from an autonomous choice. When individuals fail to develop their autonomy in the first place, responsibility lies with their social, cultural and political environment.

What does actively promoting the development of individual autonomy amount to? Answering this question is one occasion on which we affirm the place of Equal Liberty within the tradition of perfectionist liberalism. Actively promoting autonomy requires a greater degree of perfectionist intervention by the State than securing equality of freedom for autonomy development. Equal freedom for autonomy development will require public funding of performing arts centers, museums, parks, libraries, and other sources of those experiences and information. These are among the many laboratories in which experiments in living may be conducted. But access to these resources is not enough if the goal is to promote and encourage the members of a society to undertake these experiments. The use of these resources must be incentivized. And it is through incentivizing participation, over and above maintaining access, that we actively promote autonomy. An argument made by Arneson, which he takes to be an argument in favor of equality of well-being rather than the freedom to achieve well-being, is of interest on this point:

> Suppose that we could use funds from general taxation to subsidize opera (substitute some other good you deem genuinely valuable if you like) for adult citizens. We could do this in either of two ways. The first option devotes all available resources to enhancing people's capabilities (e.g. providing opportunities to see performances) and achieves a higher level of capability for all. The second option devotes some resources to propaganda aimed to persuade individuals to avail themselves of the provided opera-going opportunities. The first option provides more capability; the second secures more achieved functioning…I submit that when we imagine a clear case in which it is clear that by providing less capability for flourishing we could get more flourishing fairly distributed, one ought to opt for more flourishing. (Arneson 2000, pp. 62–63)

There is a problem with this argument which, once resolved, reveals one of the advantages of Equal Liberty over views that are concerned with the distribution of well-being itself and other views that are concerned with the distribution of freedom. The problem involves Arneson's use of the term "propaganda," which is to

say, methods of communication that exploit cognitive biases and non-cognitive emotional responses for the sake of producing a willingness to engage in the sort of activity desired by the communication's author. There is good reason to think that Arneson means what he says here, and is not merely exaggerating for emphasis. Genuine propaganda is what one would need if one were seeking a reliable, causally efficacious means to produce activity of a specific sort (such as attending the opera). But then Arneson is wrong to conclude that the propaganda-option will necessarily result in more flourishing, despite the likelihood that it will result in more participation. This is because illiberal policies like the use of propaganda also reduce flourishing, by "undermin[ing] the general character needed for excellence," the ability to achieve flourishing that is "inner and active" by recognizing the value in valuable activities and choosing them "partly for themselves rather than just as a means to something beyond them" (Hurka 1993, pp. 155, 152, 154). Hurka illustrates this point with a quote from Humboldt:

> The evil results of too extensive solicitude on the part of the state, are still more strikingly shown in the suppression of all active energy, and the necessary deterioration of the moral character…The man who is often led, easily becomes disposed willingly to sacrifice what remain of his capacity for spontaneous action. He fancies himself released from an anxiety which he sees transferred to other hands, and seems to himself to do enough then he looks to their leadership and follows it. (Humboldt 1969 cited in Hurka 1993, p. 155)

Propaganda, precisely because it relies on exploiting cognitive biases and soliciting non-cognitive emotional responses, accomplishes the "suppression of all active energy" that Humboldt describes. To grow accustomed to propaganda is to grow accustomed to being manipulated, rather than confronted with reasons which one must then respond to as a rational agent.

If we suppose, on the other hand, that Arneson does not really mean "propaganda," but instead means something like incentivizing participation in, and raising public awareness of the value of, an activity, then he runs into an equally serious problem. For such incentivizing is precisely what is required by a view, such as mine, that focuses on the distribution of capability-freedom and autonomy and is sensitive to the importance of desert. The goals of more capability and more flourishing do not come apart in this case. What Arneson's argument actually shows, then, is that a view like mine is superior to one that focuses exclusively on the distribution of access to opportunities. There are arguments against the use of public funds to subsidize cultural institutions and incentivize their use. Discussion of these objections falls under the heading of determining the appropriate limits on State action, which is the concern of Chap. 16.

I conclude the discussion of this policy goal by considering one final objection. Roemer claims that "the formation of preferences… can never be autonomous." "Preferences," he argues, "are necessarily in large part imprinted in persons from their environments, in particular from looking at the preferences of other people in their social environments" (Roemer 1994, p. 271). If this objection held up, it would in fact by detrimental to Roemer's own view. As we have seen, Roemer's theory requires that we be able to speak of the autonomous effort of individuals, and hold them responsible for the relative degree of effort they autonomously exert. But

given the Scanlonian understanding of responsibility, the autonomous formation of preferences is precisely what is required for this to be possible. Fortunately, the objection does not hold up. For it assumes that preference formation can only be autonomous if it occurs outside of any social, cultural, or institutional context, as an independent exercise in pure practical reason. But there is no reason to think that this is required. My own model of autonomous preference formation, in fact, must assume that individuals begin with a set of brute attachments, whether explained genetically or socially. What matters for autonomy is the nature of the process whereby these brute attachments are transformed into deliberative preferences; and input from other members of one's social environment is indispensable to this process.

And so my own view of distributive justice requires establishing the conditions in which individuals will be capable of developing the ability to engage in this process and encouraged to do so. The second policy goal of Equal Liberty, then, is met when two conditions are met. First, each individual is equally free to develop his autonomy. Each individual then enjoys equal access to a comparable set of varied and valuable options for engaging in experiments in living, though the option sets of different individuals may differ. The option sets are "comparable," when they are of equal cardinality, and the options that constitute them are ones that could be most preferred by similarly positioned agents as the result of well-executed courses of ends-deliberation. We determine whether this condition is met by employing the model for measuring opportunity sets developed in Chap. 7. Secondly, resources are distributed so that there is equal promotion of autonomy development. Each individual then enjoys comparable incentives to engage in these experiments in living. This is consistent with a differential distribution of resources, since we need not assume that comparable incentives can be instituted at equal cost in all communities.

The second policy goal, as should be apparent, focuses on the distribution of opportunities and encouragement for developing the rational dimension of autonomy. We have already seen that provision of the resources required to stave off frequent ego-depletion, and thus develop and exercise of a basic level of self-control, is part of the first goal. But as we saw in Chap. 7, the likelihood of self-control depends not only on preventing ego-depletion, but also on the structure of the environment in which a choice is made. Creating conditions of freedom to develop the rational dimension of autonomy as required by the second goal, however, may include employing policies which prevent choice-situations conducive to failures of self-control from emerging, and which promote choice-situations whose structures make self-control more likely in their place. The former sort of choice-situation is an unacceptable constraint on the freedom to engage in the sorts of activities crucial to developing the rational dimension of autonomy. Such policies may also be required by the third and fifth policy goals below, which concern the freedom to exercise autonomous deliberation and act on its conclusions, since the proliferation of choice-situations conducive to failures of self-control is an unacceptable constraint on these freedoms. These are a points to which we shall return in Chap. 16.

3.3 Equality of Opportunity for Autonomous Capability Choice and Development

The criteria for including an option in an agent's option set with respect to his opportunity for autonomy development are fairly weak. The agent must have sufficient opportunities to experience and learn about the option so that he can gauge his own responses to it, estimate his potential for pursuing it successfully, consider how well it fits with other options he is coming to value, understand the testimony of others regarding its virtues and benefits, etc. An agent who has developed autonomous preferences, having been given the freedom and the encouragement to do so, must then have ample opportunity to act on these preferences. He does so by autonomously choosing which capabilities for functioning to develop for himself. The next policy goal of Equal Liberty, therefore, is equal opportunity for capability development.

The criteria for including an option in an agent's option set with respect to his opportunity for capability development are more demanding. To have a capability, as we have seen, is to have the resources required for some type of valuable function plus the ability to transform those resources into that functioning. An opportunity for capability development, then, is an opportunity to develop the ability that is part of a given capability, including access to the resources required for that development. As I define such opportunities, moreover, they are opportunities to *choose* which abilities to develop, to make those choices *autonomously*, and to make those choices *from a broad range of valuable options* (where the options are, in this case, abilities one might develop). They are thus opportunities to exercise one's capacity of autonomy with respect to a range of options, and act on one's autonomous conclusions by beginning to develop one's preferred abilities. Since the agent's choice and selection of what abilities to cultivate are to be autonomous, they must be based on the agent's deliberation about what abilities will be required for realizing the conception of the good life he is developing for himself, and not be the result of threats or manipulation. To act on the conclusions one has reached autonomously is itself to exercise a capability—the capability for autonomous action. As such, it requires not only that the agent has developed the ability to choose autonomously, but also that the agent possess the resources needed to act on this choice, to transform autonomous choice into autonomous action. Included in the resources necessary to the exercise of this ability is recognition of and respect for the agent's *moral freedom* by the other members of his society. Recall the discussion in Chap. 7 of the idea that the notion of republican freedom is subsumed by the notion of capability-freedom. The reason for this was that the preference for ϕ-ing of someone with the capability to ϕ was supposed to be decisive. And this preference can only be decisive if the agent is not under a duty to refrain from ϕ-ing, if he has a right to non-interference with the exercise of his freedom to ϕ, and if he has an immunity from these rights and freedoms being altered. We can, at this point in our discussion, see that this characterization of capability was insufficiently refined. The capable agent's preference need not be absolutely decisive; his achievement of ϕ-ing must

be subject to the level of effort he exerts in exercising his capability, and to the amount of risk involved in the way he chooses to go about attempting to ϕ. But having the moral freedom to ϕ remains a necessary part of having the capability to ϕ. This concept of moral freedom will occupy us throughout Part IV.

Agents enjoy equal opportunity for capability development when they have comparable sets of valuable abilities which they may choose to develop, along with equal access to the resources required to develop those abilities. Again, the sets are comparable when they are of equal cardinality, and are made up of options which similarly positioned agents could most prefer as the result of well-executed courses of ends-deliberation. The model for measuring option sets developed in Chap. 7 thus applies to this policy goal as well. The option sets accessible to two different agents may differ, and the only capabilities that all agents must have the opportunity to develop are the capabilities required for participation in liberal political and civil society. Nor need option sets contain the options that the agents themselves would actually most prefer if they could have any set of options whatever. What is required, again, is that they contain options that could reasonably be most preferred, as the result of a well-executed course of ends-deliberation, by a similarly positioned agent—an agent of similar background, natural potential, interests, etc.

There are three objections to equality of opportunity for capability development as a policy goal. The first is as follows:

> *The Disabled: Equality of opportunity for capability development would require that we devote implausible amounts of resources to creating opportunities for the disabled, to the detriment of society as a whole.*[7]

This is the worry which led Parfit to develop prioritarianism, and remains one of the most powerful objections to maximin egalitarianism from a prioritarian perspective (Parfit 1997, p. 202). The simplest response to this objection is that it ignores the point just made about position-dependence. Nothing about my view requires that society pour resources into pursuing the impossible goal of creating an environment in which the disabled are able to develop all the capabilities of the non-disabled, or one in which those of limited natural potential are able to develop all the capabilities of those born with the potential for extraordinary mental powers or physical prowess. Differential genetic, physical and mental endowment, and thus differential natural potential for achievement, is a fact of life; and no credible theory of distributive justice can claim that we have a duty to neutralize this fact. What is required is that each agent have a comparable set of opportunities for development judged *from his own position* (or a position relevantly similar to his). Many of the capabilities which might properly be accessible to an agent of extraordinary gifts—and some of those accessible by agents of average talent and potential—would be completely out of place in the opportunity set of a disabled individual. But this does not, in itself, create a problem for my theory of distributive justice. Capabilities

[7] The very fact that this objection arises for my view, though, shows that it is a form of perfectionist liberalism which survives David Charles' first challenge. See the discussion of Charles in Chap. 9 *supra*.

which are too similar to be counted separately and distinctly in the capability set of a non-disabled individual may be appropriately counted as separate and distinct in the capability set of a disabled individual—since in the latter case, the possession of one may not naturally follow from the possession of another, as it would in the former. Once we recognize this, there is no reason to assume that establishing an equivalent degree of capability-freedom for disabled individuals—as compared, in the appropriate way, with the degree of capability-freedom enjoyed by the non-disabled—would require a ruinous commitment of resources. The goal is for the disabled individual to have the same freedom to realize his potential as the non-disabled individual has to realize his. Nothing prevents us from honestly recognizing differences both in the extent of that potential and in the appropriate way of measuring the freedom to reach it. As Raz puts it:

> Whenever we are given a choice we aspire to choose wisely, to make the best decision open to us in the circumstances. We can aspire to no less. But nor can we aspire to more. If a person does so successfully then his life is successful to the highest degree…Achievements which are beyond [one] are irrelevant to a judgment of [one's] personal well-being. (Raz 1986, p. 299)

The second objection is as follows:

> *Expensive Tastes: Some individuals would prefer to cultivate such refined abilities as the ability to discriminate the delicate floral honey notes of a Glenlivet from the robust heather honey notes of a Highland Park. But such individuals have no claim on the rest of society to keep them well stocked in single malts.*

This objection does not pose a problem for Equal Liberty. It is not a requirement of the theory that individuals be given the opportunity to develop the abilities that they would actually most prefer to develop, whatever they may be. This is because the development and exercise of those very abilities are not necessary to the agent's well-being; the background theory of well-being that I endorse does not affirm the tight connection between well-being and top-preference satisfaction that some other theories do. What is required by distributive justice is that each individual have a broad range of valuable capabilities from which to choose autonomously. An agent who has this has all he needs in order to lead a life that is, as Raz says, successful to the highest degree. He is left with nothing to complain of, even though he may lack the ability to pursue the luxurious past-times that sit atop his actual preference-ranking.

With that said, the theory is consistent with providing some individuals the opportunity to develop capabilities whose exercise requires a great deal of costly resources. This is a virtue of the view, not a vice. Biomedical research, to take one example, is an endeavor which requires access to a great deal of costly equipment. Those who are developing the ability to conduct it must have access to these costly resources during their training, as well as later on during their own independent work. The expense that must be undertaken in order to preserve the opportunity for this sort of work is justified, and not just in virtue of the fact that it is the sort of work that can properly be placed at the center of a flourishing life. A theory of distributive justice that seeks to equalize opportunities for capability development must be

sensitive to the fact that the advanced pursuit of knowledge, in a wide variety of fields, is indispensable to a society's ability to secure more valuable opportunities for all its members, both in a given generation and for generations to come. New cures and treatments for diseases and disorders are one prime example; the wide range of technological innovations that enhance the lives of the blind and deaf are another. A differential allocation of resources, including some extremely costly resources, is therefore not only consistent with Equal Liberty, but is required by it for the sake of achieving more widespread equality in the future than is possible in the present.

Finally, there is the following objection, based on one of Arneson's concerns about Nussbaum's view (Arneson 2000, p. 56):

> *The Threshold: A view that aims to maximize the number of agents who are above the threshold for opportunity/freedom/capability will have to favor those who are very close to that threshold over those who are very badly-off.*

Equal Liberty is a view that must make use of thresholds. The goal is to afford each agent a comparably *broad* range of valuable option—broad enough so that they count as free to autonomously choose what capabilities to develop. The goal is not to maximize the number of options individuals have, and in fact, the theory is perfectly consistent with the claim that at some point one has *too many* options, and is worse-off for it.[8] It is no part of my theory, however, that our primary goal should be to maximize the number of agents who do have sufficiently extensive opportunity sets. This is the feature of Nussbaum's view that makes her vulnerable to Arneson's objection, and it is not a feature my view shares with hers. The question of what social choice rule should be used to evaluate which, of any two feasible distributions of freedom, brings us closer to realizing this goal and thus should be preferred, is addressed in Sect. 3.6.2 below. This is a point which must be addressed with respect to the distribution of every level of freedom under discussion—from the freedom to develop one's autonomy, to the freedom to exercise one's capabilities—as well as with respect to the distribution of basic functioning.

3.4 Equality of Capability Subject to Effort and Capability Choice

If the first three policy goals of Equal Liberty are met, then we have a society of individuals who are autonomous, and who have equally broad ranges of opportunities to develop valuable capabilities. The capabilities which any given agent has the opportunity to develop will be ones that a similarly positioned agent—one with similar background, natural potential, and interests—could most prefer to develop

[8] This is an important point made by Gerald Dworkin, among others (Dworkin 1988, p. 65). The fact that one can have too many options is a reason not to maximin opportunity for capability development.

as the result of a well-executed course of ends-deliberation. Within the set of capabilities, any one of which an agent could choose to develop, there will be subsets of capabilities, all of whose members the agent can develop together, given a sufficient and achievable level of effort, and the resources required by the agent for that development. Let us call these subsets the agent's *feasible capability sets*. Just as a particular capability set is a set of functionings which an agent is capable of achieving, the set of feasible capability sets is the set of capability sets the agent is capable of developing.[9] If the first three policy goals of Equal Liberty are met, then, each agent will be situated to make an autonomous choice of which capability set he will develop from among all his feasible capability sets. He will do this by exercising his capacity for ends-deliberation to create a preference-ranking of feasible capability sets. The particular capabilities he chooses to develop will be the ones that enable him to achieve the functionings that constitute his conception of the good life. Of course, we need not assume that every, or even any, agent makes a choice of a whole capability set all at once. An agent may choose to work toward developing capabilities one at a time, until, over some period, he has filled in a complete feasible capability set through his choices. We do not, after all, start with a complete conception of the good life; rather, we fill one in, piece by piece, over the course of our lives.

There are a handful of capabilities that must be included in every agent's capability set. These are the capabilities to participate fully and actively in civil and political society. Just as the basic functionings are required for any valuable life whatsoever, equal capability for civil and political participation is required for life as an equal member of a liberal democratic society. And although the existence of a liberal representative democracy is by no means sufficient to ensure that the goals of Equal Liberty will be pursued and achieved—a point we will return to at the end of the chapter—it is scarcely possible that they will be in any other type of political community.[10] Having these capabilities requires that one have the opportunity to learn about, and learn the importance of, exercising them; and thus to make autono-

[9] We can imagine a feasibility constraint, set by the agent's natural potential, as something like a budget constraint. Inside the feasibility constraint are all the capability sets the agent could feasibly develop.

[10] Recently, the idea of "epistocracy"—the most moderate version of which states that those who are more knowledgeable about public policy should receive more votes—has been enjoying a good deal of popularity among some conservative political, social, and economic theorists. The irony of this is that virtually all of the beliefs of these theorists regarding economics and economic policy are *false* (as we saw in Chap. 10). The fatal flaw that permeates the epistocracy literature is the inability to recognize the relationship between the production of policy "expertise" and relations of economic power. A society that does not stifle the emergence of great disparities in wealth and economic power is likely to produce a class of policy "experts" whose judgments merely serve the interests of the wealthy and powerful, and are immune to empirical evidence and the force of the better argument (as contemporary American orthodox economics demonstrates). And just as robust democracy—by which I mean a democracy in which policy decisions reflect the most popular policy preferences—is impossible if those great disparities are not curtailed, the institution of equal and universal adult suffrage is crucial to preventing their emergence—though, again, it is by no means sufficient on its own. Liberal representative democracy requires the support of numerous specific social and economic institutions to prevent highly concentrated economic power from emerging and constraining or capturing it. See (Acemoglu et al. 2013).

mous choices about whether, when, where and how to exercise them. These opportunities, therefore, must be included in every agent's set of opportunities for autonomy development. Their actual exercise is not required. I do not assume that one must vote, for example, in order to lead a worthwhile life within a liberal society. But one must be capable of exercising them.

Equality of opportunity for capability development required that an agent have access, in one sense, to the resources he would require in order to develop any of the capabilities in his option set. The sense in which he must have access to those resources is this: should he decide to develop a particular capability, the resources required for that development would then be dedicated to him. My view does not require that the resources that would be required for the development of every capability in the set of capabilities an agent has the opportunity to develop be dedicated to that agent. The agent, in all likelihood, cannot develop all of those capabilities together. He must choose one or another of the feasible subsets. The fourth policy goal of Equal Liberty concerns the actual dedication of resources for the sake of capability development. Each agent should receive the resources he requires in order to develop the capabilities in the feasible capability set he has chosen.[11] For any given agent, then, the capability set he actually ends up with will depend only on the level of effort he autonomously chooses to exert in developing those capabilities. The fourth policy goal of Equal Liberty will have been met when this is the case.

Given sufficient resources, the only thing standing between an agent and development of his preferred feasible capability set at this point is his own effort. If we assume that every agent in the society works sufficiently and equally hard, then given sufficient resources, every agent would end up equally developing some one of his feasible capability sets.[12] If we further assume that the feasible capability sets chosen by the agents in this society are of equal cardinality, then every agent will end up equally developing an equal number of capabilities. This would be true equality of capability. Since we cannot make these two assumptions, I call the fourth goal equality of capability subject to effort and capability choice. Arneson has formulated two important objections to theories of distributive justice that are concerned primarily with the distribution of opportunities for well-being. I examine each of these objections in turn.

[11] Satisfying this condition achieves one part of what Dworkin seems to mean by "equality of resources"; the other part is achieved by satisfying the final condition of Equal Liberty.

[12] Though not *any* one of his feasible sets, as, for a given agent, different levels of effort may translate into different levels of capability development depending on the feasible capability set chosen. Individuals, moreover, are not prohibited or prevented from actually using their resources to attempt to develop capabilities which, as they have strong evidence to believe, are not feasible for them at any level of effort (a choice which would represent a failure of autonomy). The failure to develop such capabilities to a level commensurate with effort, given resources which would suffice for the development of feasible capabilities at that level of effort, is not a concern of distributive justice according to my theory.

We can express the first objection as follows (Arneson 1999a, p. 495):

> *Wanting Fewer Options: Egalitarian theories concerned with the distribution of opportunities for well-being must advocate providing an equal number of such opportunities to each agent. But an agent might wish to lead a life in pursuit of only one or two goals, and receive the resources required for those pursuits, and not have the resources relevant to those pursuits diluted for the sake of a greater number of opportunities.*

Arneson may be right that some egalitarian theories of this kind must advocate an equal distribution of opportunities to achieve well-being. But my theory does not do so, nor must it. What my theory requires is that each individual receive the resources he requires in order to develop the feasible capability set of his choosing, and that his self-development depend only on his own autonomous effort. My theory does *not* require that each agent receive the resources required to develop the same number of capabilities as any other agent is developing. The sizes of feasible capability sets may vary widely, and an agent is perfectly free to choose to develop a large one—with many fairly easily attainable capabilities, or perhaps ones that are complementary and reinforcing—or to develop a small one, in which each capability requires a great deal of time and effort to develop. Insofar as mine is a liberal theory, it must respect the freedom of autonomous agents to make choices like these.

Here is the second objection (Arneson 1999a, pp. 495–496):

> *Fetishizing Freedom: Suppose we could provide an agent with the resources he would require in order to have an additional valuable opportunity to achieve well-being. We know in advance, however, that the agent will never take advantage of this opportunity. It seems we have no reason to provide those resources. But a theory that is primarily concerned with the distribution of opportunities must advocate providing them, since it must see opportunities as valuable in themselves.*

The problem with this objection is in its assumption that a theory primarily concerned with the distribution of opportunities, or freedoms, must take them to be intrinsically, or independently, valuable. But this is false. Nor is a commitment to this view required in order to justify focusing on the distribution of opportunity. As I have already argued, focusing on opportunity—that is, on the freedom to achieve well-being, rather than on well-being itself—is justified by the need for sensitivity to the importance of desert, autonomous effort, accepted risk, and responsibility. No assumption that freedoms or opportunities are finally or independently valuable is needed. Even if we take freedom to have constitutive value, as I will argue in Chap. 16 that it does, it does not follow that a life is always improved by the addition of more freedom to achieve well-being. If an agent has already chosen the particular capabilities he prefers to cultivate from a broad range of valuable options, and we assume, with Arneson, that we know he would make no use of a further one should he be given the resources to develop and exercise it, then we do indeed lack reason to provide those resources. The additional freedom will not contribute to the goodness of the agent's life.

3.5 Equal Freedom for Capability Exercise

The fifth policy goal of Equal Liberty concerns the exercise of the capabilities that the agents in a society have developed. All individuals are to have equal freedom to exercise their capabilities in pursuit of their valuable projects and goals. Positions are to be open to all (with scarce positions open to all on a competitive basis), and are to be awarded on the impartial basis of relevant merit and qualification. We might refer to the policy required for this fifth goal as "Careers Open to Capabilities." Rawls considers a principle of "Careers Open to Talents" as a candidate for one of the basic principles of justice, but rejects it in favor of the Principle of Fair Equality of Opportunity, which I have criticized above. The reason for rejecting the latter principle was that it failed to capture genuine equality of opportunity. There is no problem, however, with endorsing a version of Rawls' weaker policy in the present context, since the first four policy goals of Equal Liberty provide just the background needed to make the policy of Careers Open to Capabilities a just one—the background that Rawls (unsuccessfully) tried to fill in when he formulated the Principle of Fair Equality of Opportunity. Careers Open to Capabilities is here just one piece in a larger puzzle. Maintaining a policy of Careers Open to Capabilities will require that the freedom of each individual to exercise his capabilities in pursuit of his goals be recognized and respected by the other members of society, just as we require that agents be free to conduct wide-ranging experiments in living as they develop their autonomy, and free to choose what capabilities to develop.[13] Assuming that equality of autonomy development presupposes equality of basic functioning, and that the Pareto and NRW Principles are satisfied, the first five policy goals of Equal Liberty are met when, for all individuals who exert equal effort and choose capability sets of equal cardinality, each of the variables in our measure of liberty from the end of Chap. 7, $L = Q_A + A(Q_D + Q_E)$, is equal.

3.6 Maximin of Achieved Functioning Subject to Effort and Accepted Risk

3.6.1 The Simple Maximin Formulation

The final policy goal of Equal Liberty is concerned with outcomes, the levels of functioning actually achieved by agents in a just society. We have learned two important lessons from our discussion of egalitarian theories of distributive justice. The first is that such theories must be sensitive to desert—which is to say, autonomous effort, the effort for whose expenditure an agent can fairly be held responsible. The ultimate objective of the policies of Equal Liberty thus far has been to articulate the conditions in which each agent receives the equal chance at a good life

[13] This is what Sen refers to as the "process aspect" of freedom (Sen 2002, p. 506).

that he deserves, and in which he can be fairly held responsible for the level of effort he puts into his own life. The second lesson is that egalitarian theories must also be sensitive to freely, knowingly accepted risk, when the agent has access to the available information which is relevant to his decision to accept the risk. In a society of equal capability subject to effort and capability choice, these two factors—the effort one exerts, and the risks one accepts—ought to be the only determinants of one's achievement of position-dependent well-being. To realize this outcome—equal achieved functioning for equal effort, subject to accepted risk—is the final policy goal of Equal Liberty. Equality of functioning is judged from the perspective of a society as a whole that endorses competitive value pluralism. Thus, any two agents that successfully exercise their capabilities to the fullest extent possible for them count as achieving equal levels of functioning.

Sensitivity to effort and sensitivity to risk cannot, to my mind, be discussed separately. Here is why. Suppose an agent possesses a feasible capability set, and goes about exercising those capabilities. Suppose he exerts an optimal level of effort in that exercise, and yet fails to achieve the level of functioning that is possible for one who is exerting that level of effort in exercising the capabilities in that set; his achievement is suboptimal. For this to be the case, some adverse happenings must have befallen him. Now suppose that we are considering a world of no risk. What this means is that every agent knows the outcomes that would result from every course of action that is open to them on every occasion. In such a scenario, this agent's sub-optimal level of achievement is of no concern for a theory of distributive justice. The agent knew in advance that he would encounter adverse circumstances, and he knew precisely how that encounter would turn out. He retains responsibility for his sub-optimal achievement despite the fact that he has exerted an optimal level of effort. There is no reason, from the perspective of distributive justice, to compensate him.

This, of course, is not the world we live in. We live in a world pervaded by risk. The relevant question, then, is always whether the risk of encountering the adverse circumstances that account for sub-optimal achievement, given the agent's level of autonomous effort, was a risk that the agent accepted in the appropriate way. Compensation for sub-optimal achievement is appropriate when the answer to this question is "No." Compensation is adequate, from the perspective of Equal Liberty, when it consists in the resources that a capable agent requires in order to move from the sub-optimal level of functioning he has achieved to the optimum achievable level of functioning, given the level of autonomous effort he has exerted.

A society in which individuals are regularly compensated for faultless sub-optimal achievement is not, I should think, the sort of society we should aspire to. The State's task in realizing the sixth policy goal is, rather, to create conditions in which agents do know what risks face them, do have access to the available information relevant to deciding whether to take those risks, and are free to choose to accept a given risk or not. Under these conditions, the amount of compensation required will be minimized (though not eliminated), and we will have the greatest attainable correlation between achieved functioning and deserved functioning. These seem like sensible goals for a just society to have. If it is just a fact about our

world that there is some risk involved in every decision and every action—if there really is no sure thing—then we must count agents as free to take a given risk or not even in cases in which there is no riskless option they could choose instead, for there will be no such options.

What remains to be specified is a social welfare function specifying how remaining resources, collected in a way consistent with the Principle of No Resource Waste, ought to be distributed once the first five goals of Equal Liberty are achieved. The final policy goal of Equal Liberty might more naturally have been equality of achieved functioning subject to effort and accepted risk. But I can see no serious objection to the claim that, with the first five goals of Equal Liberty achieved, we should, as Roemer suggests, maximin the average level of well-being achieved by the members of each effort percentile, so long as we retain sensitivity to accepted risk, even if this requires occasional departures from equality for agents who are equal in effort and accepted risk. This leaves everyone better off than they would be under a strictly equal distribution. I therefore propose the following distribution for society's remaining resources, after the first five policy goals have been met:

$$\varphi^{EL} = argsup_{\varphi} \int_0^1 \min\left[v\left(\pi, \varphi, \rho\right) \cdot \rho \right] d\pi$$

Here, the achieved well-being of the agent, measured by the valuation function $v(\cdot)$, is a function of the bundle of resources possessed by the agent (ϕ), his percentile of autonomous effort ($\pi \in \Pi = [0,1]$), and his percentile of accepted risk ($\rho \in P = [0,1]$), which, like effort, we take to be a normally distributed random variable.[14] What we are maximizing, however, is not the minimum level of achievement for each effort-percentile, but rather the minimum product of achievement and risk-percentile within each effort-percentile. We can think of this as a measure of risk-weighted achieved functioning. This is an attempt to capture sensitivity to accepted risk. For a given percentile of effort, the lowest-achieving agent may be one who took on a tremendous amount of risk. But there may be another agent whose level of achievement is only slightly higher, but who has accepted very little risk—one who fell victim to some unforeseen mishap against which he had no opportunity to insure himself. It is to the latter that available resources should be distributed. And it is the latter who will have the lower product of well-being and accepted risk, so long as his level of well-being is not *much* higher than the risk-taker's.[15]

[14] The theory assumes that achieved functioning is interpersonally level-comparable.

[15] Note that the incorporation of sensitivity to accepted risk does not expose Equal Liberty to the criticisms of "Luck Egalitarianism" advanced by Elizabeth Anderson and Samuel Scheffler. Equal Liberty does hold that individuals should be compensated for harm due to bad luck which they either could not foresee or did not have an opportunity to insure themselves against, and not for harm resulting from bad *autonomous* choices *unless that harm threatens their ability to achieve basic functioning*. But Anderson's and Scheffler's criticisms are directed against versions of Luck Egalitarianism that ignore both the moral importance of securing equality of basic functioning, and the significance of inequalities of autonomy and capability-freedom. Therefore, they do not touch the theory of Equal Liberty. See (Anderson 1999; Scheffler 2003).

What should be immediately obvious is that this closely resembles Roemer's distribution, with sensitivity to risk added in. What is different is the fact that ϕ^{EL} is not sensitive to types. This is not, of course, because in the just society individuals will cease to have features that distinguish them from one another. It is rather because in the just society, those features will not be correlated with diminished or enhanced opportunities for self-development and incentives for effort—the distribution found across the entire society will be replicated within each of the demographic groups identified in Roemer's theory, precisely because of the absence of these correlations. (If this were not the case, the prior goals of Equal Liberty would not have been achieved.) The need to recognize types in the social welfare function will be eliminated. And this goal—the goal of equalizing opportunities for development and incentives for effort across types—can plausibly be seen as the goal of Roemer's theory, considered as a transitional account. After all, the point of affirmative action, for example, is presumably to create a society in which policies like affirmative action are unnecessary. What remains, then, for a complete theory of distributive justice, is an argument that given the current state of our society—what would be specified in a descriptive account—Roemer's transitional policies are the appropriate ones to adopt with the goal of transforming society into a society of Equal Liberty. This is a task not for political philosophy, but for the combined powers of the social sciences.

3.6.2 From Maximin to Leximin

Thus far I have described the final goal of Equal Liberty in terms of the maximin social choice rule. This rule simply tells us to prefer, of any two distribution vectors, the one with the greater minimum value (and to be indifferent if these values are the same). The rule's indifference to distribution vectors with equal minima, however makes it Pareto-inefficient. So there is widespread agreement that the egalitarian who is inclined toward maximin should choose a stronger, Pareto-efficient distribution rule. The best known one, and the one I endorse, is the leximin social choice rule.[16] Leximin tells us to compare not only the minimum values, but to proceed, if these values are equal, and compare the minimum-but-one values, and prefer the distribution vector for which this is greater, etc. Given more than one ϕ^{EL} distribution, then, we should choose one based on the leximin social choice rule as our policy goal. Call this distribution $\text{Lex}(\phi^{EL})$.

Our main task in this section is to resolve the question of how to rank distributions of the goods which are the concern of the first five policy goals of Equal Liberty—from levels of achieved basic functioning to freedom for capability exercise. The rule needs to reflect a commitment to egalitarianism while simultaneously respecting the facts that functioning above a certain level no longer counts as basic, and that past a certain point having a greater number of options is not a benefit. The

[16] It is not, however, the only one. For an interesting discussion of another, see (Barbará and Jackson 1988).

answer is that we should use a *threshold* leximin social choice rule. For the first policy goal, the threshold is the level of functioning at which basic functioning has been achieved. For the others, it is the number of options which constitutes a broad range. A threshold leximin rule is just like an ordinary leximin rule, except that improvements to an individual's position above a defined threshold do not count. Let us take the case of distributions of freedom for capability choice, and assume that the two prior policy goals have been achieved. Say we currently have only two feasible distributions of this good, *A* and *B*, from which to choose. Let *n* be the threshold—the number of options which constitutes a broad range of capabilities from which to choose. Let *A* be (*n*−2, *n*−1, *n*), and let *B* be (*n*−2, *n*−1, *n*+1). According to an ordinary leximin rule, *B* is preferable to *A*; but according to a threshold leximin rule, they are equally good. Since *B* is no better than *A*, the additional resources needed to reach *B* are wasted, and so *B* is ruled out by the NRW Principle. Those resources should be used instead to further our pursuit of the next policy goal on the list. The goal of equal freedom for capability choice is of course achieved when we arrive at a distribution *C*: (*n, n, n*).

3.6.3 Other Generations, Other Nations, and Other Populations

Thus far the discussion of distributive justice has been limited to the case of the members of a particular society over the course of a single generation. And the single-society single-generation case will remain the focus in the remaining chapters. This spatio-temporal restriction is what makes the task of developing and defending a new, precise, and rigorous theory of social justice manageable. But social justice itself does not recognize these boundaries. And so if a theory developed for the single-society single-generation case is to be worth taking seriously, there needs to be some assurance that that theory suited to being expanded to address issues of intergenerational and international justice. Fortunately, there is good reason to be hopeful about Equal Liberty's prospects in both these areas. John Roemer and Roberto Veneziani have recently begun the work of reconciling the general conception of justice as equality of opportunity with intergenerational, international, and environmental concerns. The results, which I will now quickly review, could hardly be more encouraging.

First, a brief note on what might seem like a discordant note between one of Roemer and Veneziani's results and Equal Liberty. In one of their papers, they claim to show that the achievement of equality of opportunity to realize some *objective* condition, like achieved functioning, in a multi-generational society "implies the absence of human development over time" (Roemer and Veneziani 2004, p. 638). In their model, the achievement of conditions of equal opportunity for functioning implies that the average level of functioning achieved by later generations never exceeds that achieved by the first generation. What is needed to reconcile equality of opportunity with human development, they claim, is subjectivism—the view that what counts for the egalitarian, what the egalitarian tries to equalize opportunity for, is a subjective sort of *welfare*. They are careful to note that they do not claim to have

demonstrated that justice itself requires equalizing opportunity for subjective welfare; only that this must be our focus if we want a theory of social justice compatible with human development.

Roemer and Veneziani's argument, however, does not necessarily show what they claim it shows. That we must have equality of opportunity for subjective welfare if we are to have human development is certainly one valid interpretation of their result; but it is not the only one. And the other interpretation is perfectly consistent with maintaining that justice entails equality of opportunity to realize some objective condition. The other interpretation is in fact obvious from the three desiderata which Roemer and Veneziani show form an inconsistent triad: (1) protracted human development; (2) equality of opportunity for some condition; and (3) that the condition be an objective characteristic of the individual. Rather than jettison objectivism, we can, consistent with Roemer and Veneziani's result, jettison the (much less philosophically appealing) assumption of *individualism*.

A closer look at the argument will make clear what this amounts to. Their model shows that the possibility of protracted human development is compatible not with equality of opportunity for individual achieved functioning, but rather with equality of opportunity for welfare defined according to a utility function. But—and here is the crucial point—the utility function in question defines an agent's utility as the discounted sum of his own level of achieved functioning and the levels of achieved functioning of all of his descendants (Roemer and Veneziani 2004, p. 644). For an agent living at a time t, that agent's utility is:

$$u^t = \sum_{i=1}^{\infty} \beta^{i-t} F^i.$$

Here, $0<\beta<1$ is the discount factor, $i=\{1, 2, \ldots\}$ are the individuals, beginning with oneself, in one's dynasty, and F is an individual's level of achieved functioning. They then *interpret* utility in a subjectivist way, as a combination of one's own achieved functioning and the subjective satisfaction, or happiness, one derives from contemplating the achieved functioning of one's descendants.

But this is not the only available interpretation; and indeed, it is not the best one. Instead, we can interpret u as an objective but non-individualistic measure of well-being. The point would then be that one's well-being does not depend solely on characteristics of oneself; it also depends on objective characteristics of one's descendants. The measure of how good a life one has had is a function not only of what level of functioning one has achieved oneself, but also of what level of functioning they have achieved. And the reason for this is straightforward: each one of us has a tremendous influence on what our children are able to achieve, and through them, on what our grandchildren are able to achieve (and so on). The extent of one's influence on the future achievement of one's dynasty, positive or negative, is an important contributor to the goodness (or badness) of one's own life. As we project further into the future, the extent of the contribution to one's own well-being that is made by the achievements of later members of one's dynasty is lessened, owing to the discount factor. This reflects the fact that further into the future (and further

away from one's own lifetime), one's own choices and actions exert less and less influence over the level of functioning achieved by later members of one's dynasty. The radical idea here is that the value of an individual's life *cannot* be judged at the moment that life ends. It depends on the ramifications that life has on the future. It is as if we have taken Solon's dictum to "call no man happy, till he is dead" to its logical extreme; or have transposed Aristotle's conception of friendship as a relationship with "another self"—one whose good is a constituent of one's own (as are actions undertaken for the sake of his good)—onto one's relationships to one's descendants (who are more literally other selves) (Aristotle *NE* 1166b30). What is surprising about Roemer and Veneziani's result is not that it vindicates subjectivism, but that it shows that protracted human development requires that we equalize opportunity for well-being in this objective but non-individualistic sense. Human development requires that we equalize opportunity not only for individual achievement, but for individual influence on the future achievement of one's descendants. This poses no problem for my theory. There is nothing inherently individualistic about the conception of functioning I use. It can very well include success in aiding one's descendants to function well themselves. What is ultimately of interest in this result, then, is that it shows that if human development is to be possible, our concept of functioning *must* be non-individualistic in this sense, and that success in aiding one's descendants to function *must* be universally adopted as an end.

Returning their focus to well-being in the individualistic sense, Roemer and Veneziani have argued that what intergenerational justice requires is a distribution of resources which leximins opportunity for well-being across generations (Roemer 2007; Roemer and Veneziani 2007). This standard takes Rawls' Just Savings Principle, and draws out its natural implication (Rawls 1971, §44). Rawls' principle requires that each generation ensure that the next will be able to live under equally just institutions. Just institutions are governed by the Difference Principle, the most defensible interpretation of which is as a leximin distributive principle. So let us take the Just Savings Principle as requiring that each generation ensure that the members of the next generation can satisfy an intragenerational leximin distributive criterion, with the worst-off member of the next generation fairing no worse than the worst-off member of the present one. But Rawls does not give us a reason—and there does not seem to be a good one—to stop at this point, and not compare the worst-off individuals across generations, and interpret the demand of distributive justice as leximining opportunity for well-being intergenerationally as Roemer and Veneziani do.

If we model each generation as a representative agent—and this is not, I believe, a pernicious use of representative agency—then resources should be so distributed as to leximin opportunity for well-being within this set of generation-representing agents. If we assume an infinite number of generations, we will have to extend the leximin rule so that it can handle infinite-dimensional vectors. Fortunately, this can be done (Asheim and Zuber 2013). To summarize their main result: given a sufficiently high rate of technological change, a sufficiently high rate of renewal for renewable natural resources, and a sufficiently high "love of nature" among agents—where this refers to the extent to which individual well-being depends on

non-consumptive interaction with natural resources, such as the experience of hiking through pristine forest—the intergenerational leximin distribution is consistent with both environmental sustainability and positive consumption of natural resources by every generation. The relevant definition of sustainability in this context is: an intergenerational pattern of consumption of renewable natural resources which never (even given an infinite number of generations) depletes those resources down to zero. The model represents technological change as both endogenous (resulting from research and development) and costly. Unsurprisingly, the sorts of advancements represented are advancements in sustainable technology, technology whose purpose is to convert renewable natural resources into commodities with greater and greater efficiency (requiring less resources for the same level of production). Of these three factors, only one is fixed: the rate of resource renewal. Love of nature is culturally conditioned, and advances in sustainable technology are greater with greater investment. One of the crucial points the model brings out is that consumption by earlier generations which fuels advances in sustainable technology improves the position of later generations: these advances are an inherited positive externality. So the further earlier generations advance technology, the more later generations can consume consistent with environmental sustainability. That the State encourage both investment in sustainable technology and a culture that prizes nature is therefore essential to the achievement of intergenerational justice.

This result provides us with the answer to a question we have been waiting for since the beginning of the chapter. The Principle of No Resource Waste dictates that we pursue equality of opportunity for well-being within a single generation within the confines of sufficient conservation and investment in sustainable technology. We can now make precise what counts as sufficient. It is the level of conservation and investment required, given the actual rate of renewal for natural resources, to achieve intergenerational leximin. To say that each generation is constrained in its pursuit of the intragenerational leximin Equal Liberty distribution by the NRW Principle, therefore, is actually to say that the proper goal of each generation is to contribute to the overall goal of achieving the intergenerational leximin Equal Liberty distribution. Let us denote this as Intergen Lex (ϕ^{EL}).

Roemer and Veneziani are somewhat disappointed, however, when they seek to reconcile the goals of intergenerational justice and environmental sustainability with the further goal of international justice. They model a scenario in which two nations with different levels of technology—one high-efficiency, one low efficiency—share access to the planet's renewable natural resources. What they find is that, while each nation pursues the sustainability-consistent goal of an intergenerational leximin distribution of opportunity for well-being for its own residents, a divide between the well-being of those in the more advanced nation and those in the less advanced nation persists through the generations. Transfers from the more advanced nation to the less advanced that would be sufficient to close this gap are inconsistent with the more advanced nation's pursuit of intergenerational justice for its own residents. They take this result to show an irreconcilability between intergenerational justice and sustainability, on the one hand, and the *cosmopolitan* standard for international justice. According to cosmopolitanism, national boundaries

are morally arbitrary; there is a core set of moral duties which each individual owes to every other, and whether another person resides in the same geographical territory, or is a citizen of the same nation, as oneself is morally irrelevant. I am here interested in cosmopolitanism as a theory of political ethics—a view of the duties owed by States to other States or their residents. This issue is separate from the concern of cosmopolitanism as a theory of individual ethics—a view of the duties owed by one private individual to another. The nature of the charitable duties owed by individuals to other individuals is a topic I address in Chap. 14. Roemer and Veneziani take a persistent gap in opportunity for well-being across national borders as an indication that cosmopolitan justice has not been achieved (Roemer and Veneziani 2007, p. 249). But this assumes that the appropriate standard for achieving cosmopolitan justice is closing this gap completely.

We must distinguish between two types of cosmopolitan theories of international justice. Roemer and Veneziani are modeling the pursuit of what we might call *strong* cosmopolitanism, the view that what international justice requires is that wealthy States work with developing nations to eliminate differences in opportunity for well-being between the residents of different nations. But there is also *weak* cosmopolitanism, the view that what international justice requires is that each State contribute to the creation and maintenance of a global political and economic system in which no State is rendered incapable of preventing the rights of its residents from being violated, or is incentivized to itself violate the rights of its own residents.[17] As I will argue in Chap. 16, the moral authority of the individual State over its own people is grounded on the State's pursuit of the goals of Equal Liberty. The people, in turn, have the right to have their government pursue those goals. But the pursuit of social justice by one State for its people may be constrained by interference or oppression of various forms by other States. Weak cosmopolitanism recognizes a moral prohibition on such interference and oppression. It also recognizes a duty owed by one State to another to encourage the latter to pursue social justice for its own people. What weak cosmopolitanism rejects is the idea that one State can have a moral duty to bring about full-scale equality of opportunity for well-being among the population in another State.

Roemer and Veneziani's results suggest that achieving intergenerational justice and environmental sustainability is consistent with achieving the weak cosmopolitan standard for international justice. What they find is that pursuing the first two goals does not condemn each individual State to acting selfishly with respect to its strategic interactions with other States. These goals are consistent with cooperative bargaining between States (Roemer and Veneziani 2007, p. 249). And this is

[17] Weak cosmopolitanism should not be confused with (the even weaker) *negative* cosmopolitanism, a libertarian view according to which international justice only requires the satisfaction of negative duties—duties borne by States not to act in ways that harm the residents of other States. Though the world would be a great deal more just if States lived up to even this standard, I think it unlikely that satisfying only negative duties would suffice for achieving international justice. Moreover, there is no sound theoretical foundation for the libertarian view that all moral duties are negative—as will become clear in Chap. 14. These considerations largely undermine the appeal of negative cosmopolitanism.

precisely what weak cosmopolitanism requires. In particular, it requires that States form agreements that (1) prohibit State action that violates directly, or that contributes to the creation or maintenance of a global political system that violates systemically, the rights of those living under any State; and (2) coordinate the regulation of trans-national corporations, for the purpose of preventing action by corporate entities that violates directly, or contributes to the creation or maintenance of a global economic system that violates systemically, the rights of those living under any State. Satisfying this second requirement is largely a function of effectively preventing the emergence of highly concentrated global oligopoly power, which in turn depends on effectively preventing the emergence of highly concentrated national oligopoly power.[18] Since, as we have seen in the last chapter, this is crucial to economic stability, efficiency, growth and innovation, satisfying this requirement is likely to be in the interests of each State with respect to achieving intergenerational justice for its own people. And since one of the most important types of multinational corporate action to be regulated is the extraction of natural resources, satisfying this requirement is also likely to serve the goal of environmental stability.

Furthermore, weak cosmopolitanism requires that (3) the process of bargaining between developed and underdeveloped States incentivize the creation of sound political and economic institutions—the sort of institutions required to protect the rights of the individuals living under them from violation by foreign powers, and to eliminate incentives for rights violations by domestic powers—both by encouraging existing governments of underdeveloped States to enact reforms and by encouraging and empowering their people to push for them.[19] Satisfying this requirement does certainly require that agreements reached between nations include resource transfers in the form of foreign development aid; but there are relatively few forms of highly targeted aid that are effective at this stage—primarily concerned with access to vaccines, mosquito nets, nutritional supplements, clean water, fertilizer and primary education—and these are relatively inexpensive to implement (Banerjee 2007).

Satisfying the third requirement of weak cosmopolitanism is also unlikely to conflict with developed nations' pursuit of intergenerational justice for their own people. In the absence of exploitative and extractive oligopoly power, the development of poor nations yields long-term economic benefits to developed ones. It is the promise of reaping those benefits down the line that provides the incentive for developed nations to reach bargaining agreements with developing ones that satisfy the requirements of weak cosmopolitan justice. The apparent benefits to the developed world of poor countries remaining poor only manifest in a global economic

[18] This includes international cooperation to eliminate tax havens for both individuals and transnational corporations, with penalities for non-cooperative nations in the form of trade tarrifs. Tax havens currently shelter *trillions* of dollars of wealth. See (Zucman 2015).

[19] The latter is what is known as "endogenous State-building." For a discussion of actions developed nations can take to stimulate it, see "Concepts and Dilemmas of State Building in Fragile Situations: From Fragility to Resistance" (OECD Discussion Paper Series 2008).

environment in which the exercise of oligopoly power delivers an artificially low cost of living to the residents of developed nations. But developed nations themselves suffer from the existence of oligopoly power, as we saw in the previous chapter; the benefits are *merely* apparent.

The requirements of weak cosmopolitanism can also be seen as deontic constraints on each nation's pursuit of its own intergenerational distributive goal. Intergenerational justice cannot require a leximin-dominant distribution for any one nation that is only possible on account of that nation's oppression or exploitation of other nations. Any exploitative distribution is excluded from consideration, and we must understand intergenerational justice as requiring the leximin-dominant distribution which is attainable within the confines of weak cosmopolitan international justice. Roemer's account of exploitation can be used here, with nations in the place of social groups, and the global economic order in the place of the social order. Just as the advocate of Equal Liberty accepts non-exploitation as a requirement of justice in the intranational case, he must accept it in the international case. The result of international cooperative bargaining must be an agreement that is located in the core. That is, it must be an agreement that the residents of underdeveloped nations would not want to withdraw from, even if they could do so while retaining full communal control over their share of the world's physical and human capital. It must be an agreement that puts the developing nation on the path to greater prosperity than it could achieve under any circumstances (i.e. under any set of internal reforms) if it withdrew from the international economic community. The requirements of weak cosmopolitanism are meant to ensure that international agreements meet this standard. The institutional setting required to achieve weak cosmopolitan justice is likely something like the "cosmopolitan democracy" described by Daniele Archibugi, which I endorse as an essential component of an international extension of the theory of Equal Liberty. Cosmopolitan democracy calls for a set of supranational laws and legal institutions establishing a global constitutional democracy among nations, as well as allowing the peoples of the world to directly and democratically address global problems of the greatest importance (such as global anthropogenic climate change) through referenda binding on their national governments under the authority of those supranational institutions (Archibugi 2008, pp 88–122). If we accept a weak cosmopolitan theory of international justice, then, our prospects for reconciling environmental sustainability with both intergenerational and international justice are good.

To have a view of intergenerational justice and sustainability which is consistent with a weak cosmopolitan view of intergenerational justice would be a small consolation, if there were good reason to believe that strong cosmopolitanism is the theory of international justice we should adopt. The ethical attractiveness of strong cosmopolitanism, however, is largely undermined by economic and political realities. Roemer and Veneziani model the international economy with the brute assumption that there is a more advanced nation and a less advanced one; and given their purposes, they are entitled to do so. But when we turn our attention to the actual world, we cannot ignore the question of why we find more and less technologically advanced, and more and less prosperous, nations. The principle determining factors of national poverty and prosperity are (1) the nature of a nation's geopolitical

relationships with other nations, and in particular, whether or not these relationships are ones of dominance, oppression and extraction; and (2) the structure of a nation's own political institutions, and the choices made by those in power regarding the structure of economic institutions (Sachs 2006; Acemoglu and Robinson 2012). The most effective way for developed nations to positively influence the opportunity for well-being of those living in underdeveloped nations, then, is by satisfying the three requirements of weak cosmopolitanism.

The advocate of strong cosmopolitanism might acknowledge this point, but argue that should the world ever succeed in bringing about the conditions of weak cosmopolitan justice—including the condition that all nations operate under systems of political and economic institutions conducive to intranational prosperity and justice—we cannot justify stopping there, rather than going on to pursue the strong cosmopolitan standard of international justice. And in those situations in which the major institutional barriers to development have been removed, persistent gross international inequality certainly calls for further foreign aid, now targeted at the construction of infrastructure and accumulation of capital goods required for a healthy economy, in order to speed the process of development. But what the strong cosmopolitan fails to appreciate is that in such situations, it is both unnecessary and undesirable for the inequality to be minimized through foreign aid transfers alone (and thus, to be addressed through action that conflicts with a commitment to intranational intergenerational justice). The appropriate goal of foreign development aid at this stage is, as Jeffrey Sachs puts it, "to help [poor nations] onto the ladder of development, at least to gain a foothold on the bottom rung, from which they can then proceed to climb on their own" (Sachs 2006, p. 18). Just as the primary goal of intranational justice, according to Equal Liberty, is to create the social conditions in which each individual's well-being depends on his own level of autonomous effort and accepted risk, the primary goal of international justice—the weak cosmopolitan goal—is to create the global conditions in which the prosperity of each nation depends on the autonomous choices and effort of the people that make it up. Each nation could then be rightly held responsible for the level of opportunity for well-being enjoyed by its residents. In a world that achieved that goal, we should expect any residual international inequality to be small.

This does not amount to the full achievement of international justice from a strong cosmopolitan perspective. That requires an international leximin distribution of opportunity for well-being, and that distribution cannot be expected to simply fall out of the aftermath of achieving weak cosmopolitan justice. Small inequalities among nations are sure to persist, and these would likely result from differences in factors like geography and cultural attitudes. The strong cosmopolitan would insist on an obligation to eliminate these residual inequalities, even if doing so conflicts with some nations' pursuit of intergenerational justice for their own people. It is, however, exceedingly unlikely that there is any feasible way of closing the international gap in opportunity for well-being completely. Roemer and Veneziani's model represents transfers between nations as analogous to transfers between individuals—the representative agents that stand in for their generations and their governments. But interactions between sovereign nations are more complex in myriad ways. In particular, we simply cannot assume, as Roemer and Veneziani do implic-

itly, that any one sovereign nation has the same power to influence the lives of the residents of another sovereign nation as it does to influence the lives of those living within its own borders. The further goal of a leximin distribution of opportunity for well-being, which I accept in the intranational intergenerational case, cannot, in all likelihood, carry over into the international case without a global political transition to global federalism—without, that is, the creation of an entity which has the same power to influence the opportunities for prosperity of the world's nations, as a single State has to influence the opportunity for well-being of those individuals under its jurisdiction.[20]

This is the last stage in the argument against resolving the conflict between intergenerational and international justice in favor of the latter. So long as there are nation-states, achieving the strong cosmopolitan standard of international justice is likely impossible. The only way to solve the problem of international justice to the strong cosmopolitan's satisfaction entails reducing it to the problem of intranational justice, at which point the conflict with intergenerational justice disappears. I do not wish to take a stand here on whether global federalism is a good idea—or even a practicable one—or not. The point is simply that the strong cosmopolitan standard of global justice which Roemer and Veneziani use in their model, and which they find to be in conflict with intergenerational justice, is likely unachievable anyway in the real world given the existence of distinct sovereign nations. Given that, and the fact that strong global federalism would essentially reduce the problem of international justice to that of intranational justice, the choice is made for us. So long as distinct sovereign nations persist, we should pursue the goals of intergenerational justice, weak cosmopolitan international justice, and environmental sustainability.

We have good reason to be hopeful that the theory of Equal Liberty, as an intranational and single-generation theory, can be extended into an international and intergenerational theory of social justice. We also have good reason to believe that its pursuit is consistent with a commitment to environmental sustainability. And by examining the connection between sustainability and intergenerational justice we have been able to supply the NRW Principle with more precise content. Before leaving the topic of intergenerational justice, there is one more point to consider. Roemer and Veneziani do not consider intergenerational variability in population in their model, but there is no reason not to do so. We can compare the well-being of representative agents who represent populations of different sizes (Roemer and

[20] However, even if global federalism were required to achieve global justice under the strong cosmopolitan standard, it might not be required for a reasonably close approximation of it. Immediately following WWII, Keynes' grand idea for an International Currency Union—with its built-in automated mechanisms for globally redistributing trade surpluses (precisely what the current European currency union lacks, much to its detriment)—might have gone a long way toward realizing the goal of international justice, while remaining well short of full global federalism. The idea was, unfortunately, rejected by the U.S. at the Bretton-Woods conference; but there has been some renewal of interest in it. See (IMF Strategy, Policy, and Review Department Report 2010). The establishment of such a union is a likely requirement of the weak cosmopolitan standard of international justice, and I accept it as a requirement of the international extension of my theory of Equal Liberty.

Veneziani 2007, p. 229).[21] This issue is at the heart of one of the most widely discussed problems of distributive justice in the philosophical literature: the mere-addition paradox. We supposedly generate the paradox by first setting ourselves a choice between two future generations which we are in a position to bring into existence. In the first, future generation A, there are a certain number of people leading excellent lives. In the second, generation A+, there are the same number of people leading excellent lives and an equal number of people leading not excellent but still very good lives. We then get the paradox going by asserting that A+ is better than A. Now we observe that better than A+ (because it exhibits greater equality) is a future generation B in which the same number of people exist as in A+ and all are leading lives that are a bit better than very good but not quite excellent. However, if we continue this reasoning to its logical conclusion, we end up morally approving of an eventual future generation in which there is an enormous population of people, all leading lives that are barely worth living (generation Z). This is what Derek Parfit famously called the Repugnant Conclusion.

The Repugnant Conclusion has generated a large literature; but I will deal with none of it here. The key to resisting it is to diffuse the mere-addition paradox at the very first step. And the best way to do this is to uncover the source of the defective intuition that pushes us to take that first step. The crucial question is: From what perspective is generation Z, the world with the greatest possible aggregate of individual value, best? The mere-addition paradox relies on taking an external perspective, the point of view of the universe, in evaluating possible futures (I consider an objection to this assertion below). There is nothing problematic about that. But it also tacitly assumes that that point of view is occupied by a neoclassical economic agent, and considers the value of the totality of lives lived in the world both from that perspective and in that peculiar way. We can make explicit the specific perspective the paradox relies on in the following way. Suppose we personify the universe, and model it as a consumer (the Universal Consumer) and populations of the world as vectors in commodity space. A life is a commodity in a vector, and the value the Universal Consumer derives from it depends on the life's quality for the person living it (excellent, very good, etc.) The Universal Consumer, moreover, is a rational economic agent; it strictly prefers a bundle of x excellent lives plus $y<x$ very good lives to x excellent lives alone, and at a certain point is willing to trade a smaller number of higher quality lives for a greater number of lower quality lives. (From this perspective, we can account for a strict preference for A+ over A, but only a weak preference for B over A+—the greater equality of B is irrelevant, and so cannot ground a strict preference. But a weak preference here is all the paradox requires; strictly preferring B for its equality is a red herring.) The more lives we humans produce—so long as those lives are minimally worthwhile—the more value the Universal Consumer derives from the world. And so, from the Universal Consumer's perspective, the more valuable, the better, the world is. Having adopted that perspective, taking the first step in the mere-addition paradox, and then following it all the way to the Repugnant Conclusion, is as easy as applying the view that the

[21] As strange as it may sound, we say in such a case that the representative agents themselves are of "different sizes."

individual should make those consumption choices which maximize his utility, and that producers should respond to consumer demand. But once we appreciate the perspective we need to adopt in order to feel the pressure to take the first step in the paradox—the perspective of the Universal Consumer—we can see that it is a perspective we have no reason whatever to take. The argument tacitly relies on a model of the universe as a marketplace for worthwhile human lives, and the notion that only such a market-based model can provide the basis for determining what sort of world we ought to create. But that is a notion we should have discarded by now. There is no reason for us to accept that the value of the world tracks the number of people produced in it.[22]

One might object that the future version of the world in which there are people leading very good lives in addition to those leading excellent lives is better *for the former group of people*. But there is no sense in this. The thought would have to be that this world is better for them than the other world (the one in which there is only a group of people leading excellent lives) in which they do not exist. But there is no meaningful way to make this comparison. A world in which a person does not exist is not good or bad for that person in any way. It is not that such a world is of neutral value for that (non-existent in that world) person, as would be the case for a world in which the person did exist but lead a perfectly indifferent life; the value of the former world for that person is *not defined*. When we say something like "It would have been better for person X never to have been born," what we mean is simply that person X has not led, is not leading, and (likely) will not go on to lead, a life worth living. Note that this does not contradict the claim I make above, following Roemer and Veneziani, that the value of one's life may be partially determined by the values of the lives of one's descendants, even if one is no longer alive when one's descendants lead their lives. In making that claim, I assume it is possible to compare the values, for an individual, of two different generational paths, where the individual exists only during some initial segment of each path; and that those values, for the individual during the segment when he is alive, depend partially on the values of the lives of others at later segments on each path when he is no longer alive. I do *not* assume that it is meaningful to compare the values, for an individual, of two generational paths, on one of which he exists for some segment and on the other of which he does not exist at all.

We are free to assert that future generation A is better than A+ and B, and that B is better than A+. And our reason for doing so is a familiar one: the worst-off person in A is better-off than the worst-off person in B, who is better-off than the worst-off

[22] The theory which claims that what we ought to do is to create a world with the greatest possible sum of value in it is *total utilitarianism*. The fact that this view advocates a claim, the argument for which depends for its plausibility on adopting a perspective on the world akin to that of a neoclassical consumer, should come as no surprise. Classical utilitarianism developed in the same time and place as, and both influenced and was influenced by, neoclassical economics. But a commitment to single-generation utilitarianism does not commit one to total utilitarianism, and there are sophisticated versions of multi-generation, variable population utilitarianism which avoid the Repugnant Conclusion. See, for example, (Asheim and Zuber 2014).

person in A+.[23] It is idle to worry that this line of reasoning commits us to concluding that the best future generation is one in which there is a single person leading an extraordinary life. It does not. Up to a certain point, increases in population size make it possible for the members of one generation to be better-off than the members of previous, smaller generations could have been. The mere-addition paradox tacitly assumes that we have reached the point where this is no longer the case. In asserting a preference for future generation A, then, we are asserting that a future generation which is either just the same size as or smaller than the present one (depending on whether we assume the present generation is larger than A or not) is preferable, given this tacit assumption.

The size of future generations is an issue of great importance for social justice, both in absolute terms and relative to the size of the current population. But not in the way suggested by the mere-addition paradox. Let us define the growth path of a population as the generational sequence of the sizes of that population. The shape a growth path can take is restricted by the carrying capacity of the planet, which places an upper bound on population size; the maximum time a population has to reach a minimum viable size; and the limit on the rate at which a population can grow, determined by the upper bound on the fertility rate for members of the species. The opportunity for well-being possessed by the members of any given generation is in part a function of what population growth path it is on, and what place on that path it occupies. Each growth path will have its own intergenerational leximin distribution. The optimal distribution, then, is the leximin distribution on the optimal population growth path. The worst-off generation under this distribution will not only fare better than it would under any other distribution given the growth path it is on. It will also fare better than the worst-off generation under an intergenerational leximin distribution on any other population growth path. What makes the optimal population growth path optimal is just the fact that its intergenerational leximin distribution results in a worst-off generation which is better off than the worst-off generation under the intergenerational leximin distribution of every other growth path. Denote this optimal distribution as Interpop Intergen Lex(ϕ^{EL}). This would be a complete extension of the theory of Equal Liberty, and the cutting edge research of Roemer and Veneziani provides reason to be optimistic that such an extension is possible. Pursuit of this goal would require the formulation of policies which the State could enact in order to guide its population onto the optimal population growth path, without overstepping the bounds of its legitimate authority and thus violating the rights of its people, and consistent with the goal of intergenerational leximin to make the worst-off generation as well-off as possible.

[23] If the members of future generation A+ are all different from the members of future generation A, then the claim that A is better than A+ violates the person-affecting view of goodness. There is no person for whom A is better than A+; and for each life lived in A (regardless of who is living it) there is a life lived in A+ which is exactly as good. The preference-ranking A ≻ B ≻ A+ is based on the quality of life of the worst-off in each possible future generation. Accepting this criterion for judging possible future generations thus requires that we deny the person-affecting view. I am happy to do so.

4 The Social Market Economy: A Policy Program for Equal Liberty

The theory of Equal Liberty is my proposal for a theory of social justice which satisfies our three desiderata. It scrupulously avoids making the fundamental error of liberal individualism. It respects the importance of autonomous effort in determining individual well-being. And, I believe, it is non-exploitative: the society of Equal Liberty is one no section of the population would prefer to withdraw from. Our final task in this chapter is to turn to more practical matters and considers the policy implications of this vision of social justice, will serve as an argument that Equal Liberty can be achieved in a moderate social democratic form of capitalism. The practical complement to the theory of Equal Liberty is the Social Market Economy, the program of economic and social policies partially—though never fully—implemented during the years of the so-called "German miracle" of post-war economic growth and broad-based prosperity, from 1946 to 1965 (Giersch et al. 1992, ch. 2–3). In particular, I rely on the interpretation of the Social Market Economy developed by the German Catholic economist, social theorist, and policy-maker Alfred Müller-Armack—the most progressive interpretation of this idea, as well as the farthest from what was actually put into practice by Chancellors Konrad Adenauer and Ludwig Erhard, whose policies were more influenced by the ideas of the more conservative ordoliberal economic theorists Walter Eucken and Wilhelm Röpke (Glossner 2010, ch. 1.3).[24] A number of these policies, as we shall see, were also advocated in the U.S. by J. K. Galbraith. I supplement Müller-Armack's policy program with a few proposals, including some drawn from Galbraith, intended to address contemporary challenges.

The first point to make about the Social Market Economy is that it is a species of capitalism. That is to say, it is a system of socio-economic organization which recognizes private property rights, and relies on the mechanism of the price-system rather than central planning to coordinate economic activity. Müller-Armack "leaves no doubt that the social market economy is a form of capitalism," and he holds that a "widespread error was the conviction that social or social political aims were only

[24] Some readers might find it surprising to see Müller-Armack termed a progressive, or his economic theory characterized as moderately social democratic. Although Müller-Armack was not a member of the German Social Democratic Party (which was avowedly Marxist until 1959), his commitment to the idea of an active role for the State in securing the conditions of social justice (inspired by the progressive tradition within Catholic social thought) would place him, I believe, comfortably in the moderate wing of contemporary social democratic thought—in much the same way that Eisenhower's views would place him squarely within today's Democratic Party. Though he was viewed as a right-of-center thinker for much of his career, this was at a time when to be left-of-center meant to reject capitalism entirely. And though he was a member of the famous "neoliberal thought collective" the Mont Pelerin Society, the genuine neo-liberals and ordoliberals of the Society distanced themselves from him as soon as the progressive slant of his thinking became clear (Muresan 2014, p. 94). The examination of Müller-Armack's views which follows will bear this out, and substantiate the sharp distinction already drawn between neoliberalism, ordoliberalism, and the Social Market Economy.

to be achieved by switching off the rules of the market" (Koslowski 1998b, p. 77; Watrin 2000, p. 210). Nonetheless, it is guided by a different theory of capitalism from that which has dominated the English-speaking world, the theory of classical liberalism or libertarianism (which we have found to be untenable). The theory of the Social Market Economy recognizes that "as a social theory, capitalism is *materially underdetermined* and incomplete. It must be complemented by a comprehensive social-political theory concerning the framework within which capitalism can activate its advantages as a method of coordination" (Koslowski 1998b, p. 91). As Peter Koslowski explains,

> the distinguishing feature of the theory of the social market economy in contrast to other theories of the market and of capitalism [is] that it includes the institutional framework of the market in its economic theory, in its theory of the economic order. The theory of the social market economy is aware of the fact that there is not only one kind of capitalism, but a variety of capitalisms…the decision about the type of capitalism a country chooses is a question of institutional choice. (Koslowski 1998a, pp. 2–3)

The theory of the Social Market Economy, then, is a particular type of Institutional economic theory. As we will see, it assumes the same basic economic goals as the general Evolutionary-Institutional economic theory advocated in Chap. 10, and recognizes the necessity of the same traditional Keynesian countercyclical policies. But the theory of the Social Market Economy offers a specific and comprehensive vision of Institutionalist socio-economic policy that goes beyond Keynesianism, another necessity noted in Chap. 10.

The common belief that there is some deep tension between the views of Müller-Armack and those of Keynes is based on a confusion of traditional Keynesianism with the neo-Keynesian program that became dominant not only in the U.S. and the U.K., but also in Germany (in the form of the program of *Globalsteurung* which effectively supplanted the Social Market Economy) during the period of 1968–1981 (Nörr 1998, pp. 231–232). This was a program which sought to tame the business cycle completely, and used policy tools in normal economic times in an attempt to achieve pre-set macroeconomic targets—most importantly the continuous maintenance of full employment, and steady and robust annual growth in GDP (Nörr 1998, p. 233). In Germany it was, ironically, born of the Social Democrat Karl Schiller's brilliant success in using countercyclical policies, in the way advocated by both Keynes and Müller-Armack, to combat the recession of 1966–1967 (Giersch et al. 1992, p. 147). Its results following the initial periods of recovery from the oil shocks of the 1970s and 1980s, however, demonstrate that outside of recession these measures are largely counter-productive (Giersch et al. 1992, ch. 4–5). Müller-Armack's dismay with the direction of German economic policy during this period in the name of the Social Market Economy would undoubtedly have been echoed by Keynes, had he lived to see the parallel development of economic policy in the name of Keynesianism (Müller-Armack 1978, p. 9, cited in Watrin 2000, p. 211). The traditional Keynesian joins the advocate of the Social Market Economy in denying that the State performs its appropriate role "by permanently intervening in the private spheres of responsibility of entrepreneurs, employers and consumers," which interventions are "often nothing more than a sign of its weakness, a retreat in the

face of pressure from some particular economic interest" (Schlecht 1988, p. 280). The neo-Keynesian demand management of the 1970s runs counter to this ethos.

There are a number of features of the Social Market Economy which make it the ideal complement of Equal Liberty. The first is that it views the market as instrumental and basic economic goals as pursued for the sake of more fundamental social goals, just as the theory of Equal Liberty does. Müller-Armack makes this point explicitly:

> The mistakes and omissions of the [classical] liberal market economy lie in the end in the narrowness of the economic *weltanschauung* that [classical] liberalism defends. It induced [classical] liberalism to overlook the instrumental character of the order it conceptualizes and to take by mistake the market economy for an autonomous world. The market economy, however, may not at all claim to be the complete regulation of all our life…The market economy seems to us today as an instrumental means whereas [classical] liberalism has been tempted to make it the idol of it *weltanschauung*. As an instrumental means, however, the market economy continues to be for us the very efficient and up to now irreplaceable means and the way to organize the economic life of mass cultures…" (Müller-Armack 1946, translated and quoted in Koslowski 1998b, p. 79)

This recognition of the instrumental character of the market leads to another, even more significant, complementary feature between the Social Market Economy and Equal Liberty. The goal of the Social Market Economy is the achievement of a "social equilibrium":

> [T]here is a permanent ethical and political need for equalizing and equilibrating the economy in several fields. In the social market economy, the government must equilibrate between the individual economic interests in economic growth and consumption on the one hand, and the protection and preservation of the natural environment on the other hand. The state must further equilibrate between the requirements of economic freedom, efficiency, and growth on the one hand, and the need for social justice in the distribution of income and wealth and for social security on the other hand. (Koslowski 1998b, pp. 82–83)

The four socio-economic elements which the Social Market Economy seeks to balance are precisely the four elements on which the theory of Equal Liberty focuses: the market-based pursuit of basic economic goals, as a means to the promotion and preservation of individual liberty and social justice, sought within the bounds of the ecological limitations imposed by the NRW Principle.

Individual liberty is the central ideal of the Social Market Economy (Watrin 1998, p. 17). The relevant conception of liberty, however, is much broader than the one embraced by the classical liberal. According to the theory of the Social Market Economy, "the failure of the old liberals was in not recognizing that guarding private property rights and enforcing private contracts were insufficient for maintaining a liberal economic order" (Watrin 1998, p. 18). The "gross misinterpretation of the freedom of contract" of the classical liberals on the German Supreme Court during the war years "led to the creation of powerful interest groups and a new dependence of workers on their employers, of consumers on monopolists and of retailers on combines and cartels" (Watrin 1998, p. 18). The conception of individual liberty which should guide social policy is rather one that focuses on "[the individual's] independence, his opportunities for advancement and the improvement of human relations in industry." These, according to Müller-Armack, are "objectives of equal – if not far greater – importance" than the goal of "material wealth formation"

(Müller-Armack 1965, translated in Koslowski, ed. 1998b, p. 267). What individual liberty requires is "a systematically developed competitive market system," motivated by the conviction that "the rules of the competitive game of the market should be fair for everybody" (Watrin 1998, pp. 19, 18).

Müller-Armack's remarks point to equality of freedom for all individuals to exercise their capabilities—the fifth policy goal of Equal Liberty—as the central requirement of social justice. The establishment and preservation of a reasonably competitive market economy—not one which seeks to emulate the marketplace of neoclassical theory, but rather an evolutionary market economy which is free, so far as possible, from persisting, highly concentrated oligopoly power—is the essential condition for achieving equal freedom of capability exercise. Since this is a fundamental condition for economic stability, efficiency, growth and innovation, there exists a harmony between the basic economic goals and the value of individual liberty as conceived by the Social Market Economy. The theory of Equal Liberty further develops this conception of individual liberty, but recognizes that it is only one aspect of this fundamental value, and places it in the context of a comprehensive account. A market economy can only be reasonably competitive and structured in a way that is fair to all if there is equal freedom for all to develop their autonomy, autonomously choose what capabilities to develop, and then develop them. The second, third, fourth and fifth policy goals of Equal Liberty, taken together, thus provide a comprehensive account of individual liberty. The fact that a more comprehensive account of liberty is required is recognized by proponents of the Social Market Economy. And the kind of account that is required is even seen to be precisely the kind of account provided by the theory of Equal Liberty. Internal to the theory of the Social Market Economy is a recognition that more than a sound institutional theory is required to make up for the underdetermination and incompleteness of capitalism as a social theory. In addition, we need "a theory of the social genesis and normative justification of preference formation" (Koslowski 1998b, p. 91). The theory of Equal Liberty is the only theory of social justice that properly addresses this need. Its account of individual liberty is built on the theory of normatively justified preference formation developed in Chaps. 3 and 4, and it describes the social conditions under which the formation of justified preferences can best be encouraged.

The essential precondition for achieving the second, third and fourth goals of Equal Liberty is likewise the creation and preservation of a reasonably competitive market economy, in which oligopoly power is significantly curtailed. We have already noted that the development and exercise of autonomy is most effectively encouraged by a competitively pluralistic society. The competition referred to in that context is that between different systems of values. But genuine and robust competitive value pluralism is only possible in the context of a reasonably competitive market economy. Oligopoly power is the enemy of both value pluralism and autonomy. The sources of conflict between value pluralism and the encouragement of autonomy development on the one hand, and the existence of oligopoly power on the other, is apparent from the previous chapter: oligopolies exercise a disproportionate influence over public policy geared to their own ends, and a disproportionate amount of persuasive power over the individual consumer's decisions. The exercise

of these two powers has a profound effect on the value system that characterizes a society and its culture. Specifically, it creates a value system which is shaped by the belief that, as Galbraith puts it,

> the purposes of the planning system are those of the individual…that any public or private action that serves its purposes serves also the purposes of the public at large…that the production and consumption of goods, notably those provided by the planning system, are co-ordinate with happiness and virtuous behavior. Then all else becomes subordinate, more or less, to this end. (Galbraith 1973, p. 241)

In encouraging the development and exercise of autonomy, the State must work to curtail the emergence of oligopoly power. And so it must work to create and preserve a reasonably competitive market economy. This is one reason why the second, third and fourth goals of Equal Liberty exist in the same harmony with the basic economic goals as the fifth does. Another is that the society that achieves these goals makes the broadest possible investment in its human capital, one of the most crucial prerequisites for stable, sustainable long-term economic growth. As Joseph Stiglitz has recently argued, the lack of this kind of investment in the capabilities of their own members is one of the primary reasons why severely unequal societies tend to exhibit such destructive instability in the long run (Stiglitz 2012, p. 108). This argument is supported by the evidence presented in recent reports from the IMF and the OECD.[25]

One might object that a liberal, perfectionist, egalitarian theory—like the theory of Equal Liberty—should not be concerned with preserving competitive markets. Galston discusses this objection, which he attributes to John Schaar: "[Competition] is a defective mode of existence. It sets human being apart from each other and pits them against one another, in an essentially destructive struggle" (Schaar 1967 cited in Galston 1991, p. 208). Galston provides a good response to this objection. It is true that any form of competition, even the good kinds that Galston identifies, will involve disappointment and frustration on the part of some agents on some occasions. If there is competition, we will not all succeed at everything all the time. This, however, does *not* count against the value of competition from the perspective of a perfectionist liberalism. As Raz has observed, "anguish, frustration, and even suffering are often part and parcel of rewarding activities and experiences, which depend [in part] on the suffering, etc., for their meaning, and therefore for their value as well…frustration and anxiety, or at least the ever-present risk of them, are common elements in most of the relationships, activities, and undertakings of human life…" (Raz 1994, p. 19).

Thomas Hurka argues that the liberal perfectionist egalitarian must claim that

> perfection is not entirely *competitive*. If it were always a necessary condition of one person's reaching a level of perfection that others do not, there could never be a universal advance in excellence, and there would be no point in pursuing such an advance by distributing resources equally. (Hurka 1993, p. 176)

Hurka then argues for the stronger claim that perfection is co-operative, since the development of one member of a society is encouraged by, and encourages, the development of others (Hurka 1993, p. 176). Among the arguments he gives for this

[25] See the work of Ostry and Berg cited *supra*.

claim is one that invokes the value of collaborative activity (Hurka 1993, pp. 177–178). Nothing I have said thus far regarding the value of promoting and preserving reasonable levels of competition within markets, or the importance of endorsing competitive value pluralism, is at odds with these claims. There is ample room within my theory to recognize the value of co-operation and collaboration. Extensive co-operation is required for creating the conditions in which each of the policy goals of Equal Liberty are achieved. Müller-Armack was a proponent of "governmental efforts to increase the elements of co-operative self-support in the economy" (Koslowski 1998b, p. 86). The duty of each individual to contribute to this project will be examined in Chap. 16. The fact of competitive value pluralism, moreover, does not count against the possibility or the value of co-operation. The most effective collaborations often take place among individuals whose lives are structured by very different values. It is those different values that lead them to develop a wide range of abilities that no one person could master. So long as there is mutual respect among participants, it is the combination of those different abilities that produces the most productive and innovative collaborative enterprises.[26] Finally, we must remember that the theories of Equal Liberty and of the Social Market Economy are not, and are not meant to be, strong ideal theories. Both are consistent with the perspective of modern Catholic social thought, according to which the best system of social organization is that which would emerge if every person could be relied on to follow the dictum to "love everybody the same as ourselves…the market [is] the second-best solution in an ideal world but [is] the best possible solution under conditions as they are, under the human condition as it is" (Koslowski 1998a, p. 9).

In addition to individual liberty, the theory of the Social Market Economy recognizes social security as a second fundamental social value. It is not "a fundamentally alien appendage to the system," but "an essential integral component," which is a prerequisite of any acceptable economic system (Schlecht 1988, p. 282). This commitment is represented in the theory of Equal Liberty by the priority of the goal of universal basic functioning. The relationship between this goal and the basic economic goals, however, is "not one that is completely free of friction" (Schlecht 1988, p. 282). The provision of social assistance is, in many cases, a form of investment for the future: those who are aided are thereby enabled to become, or return to being, productive participants in the market economy who require no assistance at a later time. But this is certainly not true in all cases. Some individuals never will, or never again will, find themselves in that fortunate position. But the Social Market Economy recognizes it as a social duty to enable "above all, the weakest members of society to lead a life befitting a human being," and trusts that "in a market economy the strong incentives to produce would overcompensate the welfare losses caused by social policy" (Watrin 1998, pp. 19, 21). The basic economic goals, recall, have only an instrumental priority. We pursue them, among other reasons, because "[w]ithout efficient economic management it is impossible to raise the enormous level of funding required to provide a dignified and socially acceptable livelihood for those who, under market conditions, are not in a position to earn their

[26] For an excellent historical study of a prominent instance of this phenomenon and its importance, see (Goodwin 2005).

own living, be it on a temporary basis or in the longer term" (Schlecht 1988, p. 282). Some degree of sacrifice of economic efficiency and growth potential must be accepted, so long as it is not so great as to endanger the State's ability to continue to pursue the goals of Equal Liberty for future generations. The proponent of the Social Market economy recognizes, however, that "[i]t might become necessary to sacrifice parts of a social security system that have become obsolete, wasteful or simply too expensive and expansive." How exactly that balance is to be struck is a question we will come to when we examine specific economic policy recommendations and consider how they may best be interpreted.

The third fundamental value in the Social Market Economy's social equilibrium is environmental preservation. This commitment is represented in the theory of Equal Liberty by the role of the NRW Principle. The economist Otto Schlecht posits as a fundamental thesis of the theory of the Social Market Economy that "[p]rotecting God's own creation by securing our natural environment is one of the central ethical postulates, if not the greatest challenge facing the responsible individual and society in the industrial age" (Schlecht 1988, p. 285). The proponent of the Social Market Economy is moved by a conviction that "[p]rovided the market and competition are functioning correctly—and provided there is an appropriate [institutional] framework—producers and consumers can adapt in a rapid and comprehensive manner to new ecological constraints and demands" (Schlecht 1988, p. 286). In particular, Schlecht recognizes the importance of developing new and ever more efficient sustainable technologies—one of the most significant factors in Roemer and Veneziani's results on intergenerational justice and sustainability. This adaptation "does not mean the emphasis can be placed on the subsequent cleanup of environmental damage, but rather that production processes and consumer decisions are altered so that environmental damage is in fact avoided in the first place. Private initiative in an atmosphere of competition promotes technological development and innovation, in the environmental sector as elsewhere" (Schlecht 1988, p. 286).

5 Conclusion

The theory of the Social Market Economy has a strong claim to be the appropriate practical complement to the theory of Equal Liberty. It is an Institutionalist theory that views the market as instrumental, takes a similar conception of individual liberty as its central value, and seeks an equilibrium with the values of basic social security and environmental preservation. Our theory of social justice, in both its theoretical and (at an admittedly very general level) practical aspects, is now complete. Our final task is to argue that the State has the moral authority and the moral duty to pursue the goals of Equal Liberty, and that the bounds of legitimate State action are delimited by this pursuit. This task is motivated by one final conviction of proponents of the Social Market Economy: "for a state to claim moral legitimacy, it must face up to its responsibility by solving those social tasks that cannot be solved by the market and competition in a socially acceptable manner…" (Schlecht 1988, p. 280). Defending this conviction will occupy us for the remainder of the book.

References

Acemoglu, D., and J.A. Robinson. 2012. *Why nations fail: The origins of power, prosperity, and poverty*. New York: Crown.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J.A. 2013. Democracy, redistribution, and inequality. NBER Working Paper No. 19746. An Act to Study and Modify Certain Coastal Management Policies. 2012.

Anderson, E. 1999. What is the point of equality? *Ethics* 109: 287–337.

Archibugi, D. 2008. *The global commonwealth of nations: Toward cosmopolitan democracy*. Princeton: Princeton University Press.

Aristotle. *Nicomachean ethics* [*Ethica Nicomachea*]. Oxford Classical Texts, ed. L Bywater. Oxford: Oxford University Press.

Arneson, R. 1999. Equality of opportunity defended and recanted. *Journal of Political Philosophy* 7(4): 488–497.

Arneson, R. 2000. Perfectionism and politics. *Ethics* 111(1): 37–63.

Asheim, G.B., and S. Zuber. 2013. A complete and strongly anonymous leximin relation on infinite streams. *Social Choice and Welfare* 41: 819–834.

Asheim, G.B., and S. Zuber. 2014. Escaping the repugnant conclusion: Rank-discounted utilitarianism with variable population. *Theoretical Economics* 9: 629–650.

Banerjee, A.V. 2007. *Making aid work*. Cambridge, MA: MIT Press.

Barbará, S., and M. Jackson. 1988. Maximin, leximin, and the protective criterion: Characterizations and comparisons. *Journal of Economic Theory* 46: 34–44.

Benn, S. 1988. *A theory of freedom*. Cambridge: Cambridge University Press.

Dworkin, G. 1988. *The nature and value of autonomy*. Cambridge: Cambridge University Press.

Galbraith, J.K. 1973. *Economics and the public purpose*. New York: Pelican.

Galston, W.A. 1991. *Liberal purposes*. Cambridge: Cambridge University Press.

Giersch, H., K.-H. Paqué, and H. Schmieding. 1992. *The fading miracle: Four decades of market economy in Germany*. Cambridge: Cambridge Universtiy Press.

Gilens, M., and B.I. Page. 2014. Testing theories of American politics: Elites, interest groups, and average citizens. *Perspectives on Politics* 12(3): 564–581.

Glossner, C.L. 2010. *The making of the post-war German economy*. London: I.B. Tauris.

Goodwin, D.K. 2005. *Team of rivals: The political genius of Abraham Lincoln*. New York: Simon & Schuster.

Hayes, T.J. 2013. Responsiveness in an era of inequality: The case of the US Senate. *Political Research Quarterly* 66: 585–599.

Hurka, T. 1993. *Perfectionism*. Oxford: Oxford University Press.

IMF. 2010. *Reserve accumulation and international monetary stability*. IMF Strategy, Policy, and Review Department Report. Washington, DC: IMF.

Jacobs, A.M. 2011. *Governing for the long term: Democracy and the politics of investment*. Cambridge: Cambridge University Press.

Kenny, C., and A. Kenny. 2006. *Life, liberty, and the pursuit of utility: Happiness in philosophical and economic thought*. St. Andrew's: St. Andrew's Press.

Koslowski, P. 1998a. The social market economy and the varieties of capitalism: Introduction. In *The social market economy: Theory and ethics of the economic order*, ed. P. Koslowski, 1–12. Berlin: Springer.

Koslowski, P. 1998b. The social market economy: Notes on Müller-Armack. In *The social market economy: Theory and ethics of the economic order*, ed. P. Koslowski, 73–95. Berlin: Springer.

Lerner, A. 1944. *The economics of control*. New York: MacMillan.

Müller-Armack, A. 1946. *Wirtschaftslenkung und Marktwirtschaft*. Hamburg: Verlag für Wirtschaft und Sozialpolitik.

Müller-Armack, A. 1965. The principles of the social market economy. Trans. Peter Koslowski. In *The social market economy: Theory and ethics of the economic order,* ed. P. Koslowski, 255–274. Berlin: Springer.

Müller-Armack, A. 1978. Die fünf großen Themen der künftigen Wirtschaftspolitik. *Wirtschaftspolitische Chronik* 27: 1.
Muresan, S.S. 2014. *Social market economy: The case of Germany*. Dordrecht: Springer.
Nedelsky, J. 2011. *Law's relations: A relational theory of self, autonomy and law*. New York: Oxford University Press.
Nörr, K.W. 1998. Social market economy and legal order in Germany. In *The social market economy: Theory and ethics of the economic order*, ed. P. Koslowski, 220–247. Berlin: Springer.
OECD. 2008. *Concepts and dilemmas of state building in fragile situations: From fragility to resistance*. Paris: OECD.
Parfit, D. 1997. Equality and priority. *Ratio* 10(3): 202–221.
Rawls, J. 1971. *A theory of justice*. Cambridge, MA: Harvard University Press.
Rawls, J. 1987. *Political liberalism*. Cambridge, MA: Harvard University Press.
Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.
Raz, J. 1994. Duties of well-being in ethics. In *Ethics in the public domain: Essays in the morality of law and politics*, 3–28. Oxford: Clarendon Press.
Roemer, J. 1994. *Theories of distributive justice*. Cambridge, MA: Harvard University Press.
Roemer, J. 2007. Intergenerational justice and sustainability under the leximin ethic. In *Intergenerational equity and sustainability*, ed. J. Roemer and K. Suzumura, 203–227. New York: Palgrave MacMillan.
Roemer, J., and R. Veneziani. 2007. Intergenerational justice, international relations, and sustainability. In *Intergenerational equity and sustainability*, ed. J. Roemer and K. Suzumura, 228–251. London: Palgrave MacMillan.
Roemer, J., and R. Veneziani. 2004. What we owe our children, they their children…. *Journal of Public Economic Theory* 6(5): 637–654.
Sachs, J. 2006. *The end of poverty: Economic possibilities for our times*. New York: Penguin.
Schaar, J. 1967. Equality of opportunity and beyond. In *Nomos 9: Equality*, ed. J.R. Pennock and J.W. Chapman. New York: Atherton.
Scheffler, S. 2003. What is egalitarianism? *Philosophy and Public Affairs* 31: 5–39.
Schlecht, O. 1988. The ethical content of the social market economy. Trans. Peter Koslowski. In *The social market economy: Theory and ethics of the economic order,* ed. P. Koslowski, 275–289. Berlin: Springer.
Sen, A. 1996. Rights: Formulation and consequences. *Analyse und Kritik* 18: 153–170.
Sen, A. 2002. *Rationality and freedom*. Cambridge, MA: Harvard University Press.
Stiglitz, J. 2012. *The price of inequality: How today's divided society endangers our future*. New York: WW Norton.
Tuggoden, B., and P. Vallentyne. 2010. On the possibility of a paretian egalitarianism. *Journal of Philosophy* 102(3): 126–154.
UNEP. 2008. *The economics of ecosystems and biodiversity: An interim report*. Nairobi: UNEP.
UNEP Finance Initiative. 2010. *Universal ownership: Why environment externalities matter to institutional investors*. Nairobi: UNEP.
von Humboldt, W. 1969. *The limits of state action*. Trans. J.W. Burrow. Cambridge: Cambridge University Press
Watanabe, K., et al. 2005. Early childhood development interventions and cognitive development of young children in rural Vietnam. *The Journal of Nutrition* 135: 1918–1925.
Watrin, C. 1998. The social market economy: The main ideas and their influence on economic policy. In *The social market economy: Theory and ethics of the economic order*, ed. P. Koslowski, 13–28. Berlin: Springer.
Watrin, C. 2000. Alfred Müller-Armack: Economic policy maker. In *The theory of capitalism in the German economic tradition*, ed. P. Koslowski, 192–220. Berlin: Springer.
Zucman, G. 2015. *The hidden wealth of nations: The scourge of tax havens*. Chicago: University of Chicago Press.

Part IV
Justice: Authority

Chapter 12
Justice: Authority – Introduction

We have been occupied with the capacity of autonomy, and the freedom to develop and exercise that capacity and to act on the conclusions reached through its exercise, for quite some time now. As we shift our attention from the question of what the State's distributive ideal should be to the question of what grounds the State's authority to pursue a scheme of distributive justice, and what limits on that authority should be in place, we must begin to consider a distinct but related type of freedom, which I will call "moral freedom." An agent is morally free to perform an action ϕ if, and only if, that agent is not under a moral duty to refrain from ϕ-ing. The following chapters include a detailed discussion of the nature of moral duty and of the features that make it the case that one agent has a moral duty to another, including the moral duty to comply with the directives of a practical authority. In this introductory chapter, I limit myself to three brief tasks. First, I will say a few words about the relationship between autonomy-freedom and moral freedom. Second, I will provide a short introduction to the Hohfeldian analysis of rights. And third, I will discuss the relationship between the Hohfeldian formulation of rights and the social choice-theoretic formulation of rights. Although the apparatus and methods of social choice theory have thus far proved invaluable in formulating a precise characterization of autonomy-freedom, my final two chapters will deal with rights solely from a Hohfeldian perspective. Nonetheless, the social choice-theoretic approach to rights complements the Hohfeldian approach in an important way, by providing a way of evaluating whether an individual's purported rights are actually being recognized and respected by his society.

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_12

1 Autonomy-Freedom and Moral Freedom

In assessing the extent of an agent's autonomy-freedom, we focused on the range of valuable options available to him. A valuable option was taken to be one which might be most preferred by some agent as the result of a well-executed course of ends-deliberation. The options that count toward the measure of an agent's freedom should be restricted to those that the agent is morally free to pursue—those options which *can* be pursued by the agent *without* the agent having to perform any action which he is under a moral duty to refrain from performing. The question is whether this actually constitutes an additional restriction, or whether it is implied by the method for determining the extent of an agent's freedom that is already in place. For as I discussed in Chap. 4, a well-executed course of ends-deliberation takes into account considerations such as the categorical reasons that the agent recognizes as applying to him, and the agent's judgments regarding the impact of his pursuit of his available options on the lives of others. So we have some good reasons to think that the options that end up counting toward the extent of the agent's autonomy-freedom will simply end up being options which the agent is also morally free to pursue. Nevertheless, I do not wish to fully commit myself to this claim. The types of considerations just mentioned are not definitive, and we may want to leave open the possibility that they are overcome by other types of considerations even in a well-executed deliberation—perhaps an agent's emotional engagement with a particular pursuit is exceptionally strong, the goal's fit with his other goals is particularly tight, and he is encouraged by misguided or misleading, but still very convincing, testimony. I am willing, therefore, to accept a requirement of moral freedom as an additional restriction on the range of options that can count toward an agent's autonomy-freedom. In addition to the features of a situation that make it the case that an agent is (or is not) morally free to take a course of action, the relationship between ends-deliberation and ethical deliberation, and thus between autonomy-freedom and moral freedom, will receive careful attention in the next chapter.

1.1 The Hohfeldian Analysis of Rights

The standard taxonomy of rights is due to the jurist Wesley Newcombe Hohfeld (Hohfeld 1946):

Claims: A has a claim against B that B ϕ iff B has a duty to A to ϕ.
Liberties: A has a liberty to ϕ iff A is not under a duty to refrain from ϕ -ing.
Powers: A has a power over B iff A has the ability to confer a claim-right or a liberty on B, or to strip B of a claim-right or a liberty, or to confer on B the ability to do likewise to C, or to strip B of that ability.
Immunities: A has an immunity against B iff B lacks some power over A.

Hohfeld intended his analysis to apply to legal rights, but it may be understood as an analysis of either legal or moral rights. Moral claim-rights and moral liberty-rights, on this understanding, are defined in terms of moral duties. Powers, from a moral perspective, are normative powers: they are abilities to act in such a way that one affects what reasons for action are had by other agents. A moral claim-right of one agent, for example, may be created or extinguished through actions which affect the moral duties of another agent. And any action which affects what moral duties an agent has will thereby affect what moral liberty-rights he possesses. In Chap. 14, I discuss the incomplete nature of the simple Hohfeldian characterization of claim-rights, at least in the case of moral claim-rights. In Chap. 15, I discuss morally legitimate practical authority as a combination of a specific type of normative power with a claim-right held against those subject to the authority. Chap. 16 explores the sources of the limitations of the exercise of legitimate authority—that is, the sources of the moral immunities retained even by those who are subject to a legitimate authority. But for now, these basic definitions will suffice. My final task here is to relate the Hohfeldian characterization of rights to the social-choice theoretic formulation which allows us to evaluate whether an agent's exercise of his rights has the appropriate effect within the context of his society.

1.1.1 Social Choice Theory and the Structure of Rights

Amartya Sen has offered the following social choice-theoretic axiom as a plausible requirement on any society that claims to respect its members' freedom of choice:

> Minimal Liberty (ML): Each individual in a society should have a recognized personal sphere in which his preferences and his alone would count in determining the social preference. Formally, for all individuals i in a society S, and for all outcome-types x,y within a choice set C_{ps} constituting the individual's recognized personal sphere, $xP_S\, y$ iff $xP_i\, y$ (Sen 2002, p. 384).[1]

The term "social preference" requires some explanation. Though a number of interpretations are possible, I will take social preference to indicate the outcomes that are actually obtained within a society. Suppose, for example, that one member of a society would like to read a particular novel (as in Sen's famous *Lady Chatterly's Lover* argument). Suppose further that he has the ability to read the novel (he is literate), he has time to read it, he has access to a bookstore and enough money to purchase it, etc. In short, he has the capability to satisfy his preference. Other members of the society, however, would prefer it if no one in their society were permitted to possess and read this novel (perhaps because they find its content objectionable). If the actual outcome which obtains in this society is that the one who would like to read the novel does not end up reading it—he is prohibited from acquiring it and

[1] Sen's requirement ML actually requires only that at least two members of a society have such a personal sphere consisting of at least two possible options. This very weak assumption is all that is required in order to generate a requirement with the Pareto principle. I will discuss the stronger version of ML, according to which every member of a society must have such a personal sphere.

others are prohibited from furnishing him with it—then the social preference is that he not read the novel. We may then interpret requirement ML as follows. The way in which decisions are made in a society should be so organized that each member of that society has a recognized personal sphere within which his preferences are decisive. That is to say, society should be organized in such a way that each individual is left perfectly unobstructed in his exercise of his capabilities for the sake of satisfying certain of his preferences. In these cases the social preference—the outcome that actually obtains—will match the individual's preference—the outcome which the individual prefers.

The point of requirement ML is to capture a plausible and intuitive understanding of the right to self-determination of a member of a free society. In some cases—those that fall within a recognized personal sphere—an individual should be the sole determinant of what happens. We have, to stay with the example, a right to read the novels we would like to read, and this right ought to be respected by society. But ML does not specify either a particular right, or type of right, or even the form of a right that we believe individuals have and that ought to be respected. Rather, it specifies the relation which holds between an individual's preferences and particular outcomes in those cases in which the individual's right to self-determination is respected, and it asserts that in a free society there ought to be some such cases. It provides us with a criterion of success for respecting individuals' rights. But what is it precisely that we must recognize in order for ML to be satisfied? What is the structure of the sort of right we must respect in order for the ML relation between preferences and outcomes to obtain? And how must society be organized so that individual preferences within a recognized personal sphere are indeed decisive?

We can answer these questions from a Hohfeldian perspective. Let us begin by assuming that we are able to identify the options which ought to be included in an agent's personal sphere (we will be able to jettison this assumption shortly). Our focus for the moment will be on the legal rights which must be recognized and respected in order for ML to be satisfied. First, and fairly obvious, is that it must be the case that the agent has a liberty-right to act on his preference for x over y (for reading the novel over not reading it, say), by choosing x (reading the novel). But a great deal more than this is required. The agent must have a claim-right against every other member of the society that they not interfere with his reading the novel (with appropriate qualifications—that they not interfere with his reading during his own leisure time, say). Now suppose that someone does interfere—by spotting him reading this novel in the park and confiscating it from him. There must be someone with the power to impose on the confiscator the duty to return the book, and to enforce that duty. The agent must be at liberty to petition this person to exercise that power. And, we might further insist, the one who has this power must have a duty to exercise it in this way, once he has determined that our agent's claim right has been violated. And perhaps the agent must, in addition, have immunities protecting his liberty to read the novel, his claim-right against interference by other individuals, and his liberty to seek redress should that claim-right be violated.

The Hohfeldian and social choice-theoretic approaches to rights, then, are complimentary in an important way. Social choice theory specifies what ought to happen when society is organized so that the freedoms of its members are properly recognized and respected. Hohfeld's analytical framework then allows us to sketch the structure of the required social organization. This framework also enables us to answer the question of what determines the appropriate contents of the personal sphere. They are those options which the agent is *morally* at liberty to choose, and concerning which the agent has a moral claim-right of non-interference, a right which a legitimate authority would have the normative power to protect and enforce, etc. The purpose of the following chapters is to explore the grounds for the moral rights, freedoms, and duties of individuals within a society, including the duty to comply with legitimate authoritative directives, as well as the ground and appropriate limitations of the authority to create new moral duties and to enforce existing moral rights and duties. The theory that emerges will provide a cogent account of the moral basis for a scheme of social organization that respects the moral rights and freedoms of individuals within the context of establishing equality of liberty.

References

Hohfeld, W.N. 1946. *Fundamental legal conceptions as applied in judicial reasoning*. New Haven: Yale University Press.

Sen, A. 2002. *Rationality and freedom*. Cambridge, MA: Harvard University Press.

Chapter 13
Moral Reasons and Moral Duties

1 Introduction

My task in this chapter and the next is to develop a theory which explains the existence of moral rights. I will need to have a theory of moral rights in place in order to argue, in the chapters that follow, that the members of a political community have a moral right that the State pursue the goals of the policies of Equal Liberty, and that a State that adopts the policies of Equal Liberty has a right to compliance with its directives, insofar as those directives are necessary to achieving those goals. Such a theory must identify the facts, and the features of persons, in virtue of which persons possess moral rights, and explain how those facts and features give rise to moral rights. Speaking broadly, two types of accounts have been offered in the philosophical literature: status theories and teleological theories. Briefly, a status theory claims that an individual possesses moral rights in virtue of some of her essential properties, often her rationality, autonomy and dignity. A teleological theory claims that an individual possesses a moral right in virtue of something valuable that is likely to be promoted by her possessing and exercising that right. The sort of teleological theory I am concerned with claims that a right-holder possesses her rights in virtue of her interests, which her rights protect and promote.

I aim to develop and defend a new teleological theory, one which is able to meet the challenge of justifying the enforcement of rights. All acts of coercion require justification. Sometimes the fact that the good that will result from the coerced act outweighs the harm to the person coerced is sufficient. But the justification for enforcing a right must be more complex than this. One way in which we can be harmed by coercion is by being deprived of the freedom to determine our own actions. That one person has a right, however, implies that another is under a duty, and thus is not at liberty to refrain from satisfying the interest protected by the right. The justification for enforcing a right, then, must demonstrate that the violator lacks this freedom in this particular case, rather than showing that the harm of violating this freedom is outweighed by some other good. Any theory that explains why individuals have rights must explain why the right-holder and/or members of her

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_13

community have this particular sort of justification for enforcing her rights. This requirement follows from the fact that rights just are the sort of thing whose enforcement is justified in this way; that is part of what needs to be explained. Let us call this the justification constraint on theories of rights.

Contemporary status theories have a way of meeting the constraint, but it involves positing a hypothetical, but nonetheless binding, contract between all persons considered purely as rational autonomous agents. In brief, these views assert that all persons would, if they were acting purely as rational and autonomous agents, bind themselves in a communal agreement to recognize certain rights of all members of the moral community; and since we all should so bind ourselves, insofar as we should act as rational and autonomous agents would, we are all bound to abide by the substance of such a hypothetical agreement, despite not having actually entered into it.[1] I take their reliance on such an exotic theoretical apparatus to be a sound reason to reject these views, assuming that a viable alternative is available. Contemporary teleological theories like Joseph Raz's, which I discuss in detail below, cannot meet the constraint. Such theories make two claims: first, that individuals possess rights in virtue of the great value of satisfying the interests which those rights would protect; and second, that the duties correlative with rights are grounded on those rights. I argue that theories of this type fail because we can only derive the right sort of justification for coercion from interests if those interests themselves ground duties. Since giving this justification is part of explaining the existence of rights, grounding duties on interests is one step in the explanation of why rights exist: duties are explanatorily prior to rights.[2] I thus offer a new teleological theory in which duties are grounded on individuals' interests, and individuals' rights exist in virtue of the duties owed to them. The present chapter focuses on the nature of moral duty. The next chapter shows how moral rights emerge from moral duties of a specific kind, and argues that my theory satisfies the justification constraint, and so gives at least as good a justification of enforcing rights as a status theorist without having to posit any hypothetical social contract.

2 The Problem of Moral Duty

Since my theory of moral rights will contend that such rights exist in virtue of the existence of moral duties which are owed to the purported right-holders, I must give an account of the nature and sources of moral duties. This task leads me to grapple with one of the great challenges in modern ethical theory.

[1] For a recent well-developed example, see (Darwall 2006).

[2] For a discussion of the tradition of prioritizing rights, not limited to teleological theories, that traces this view back to Locke, see (Raz 1994, pp. 29–30). There is a competing tradition that prioritizes duties, which is prominent in the writings of Bentham and Austin, and comes to fruition in Mill. As I explain below, Raz's theory and my own can be seen as alternate ways of developing the welfarist theory of rights developed by JS Mill in *On Liberty* and *Utilitarianism*, without subscribing to Mill's utilitarian account of right action. For a good discussion of Mill's theory, see (Sumner 1987, pp. 132–142).

2.1 Anscombe's Challenge

Over 50 years ago, Elizabeth Anscombe identified what is probably the major challenge for non-theistic theories of ethics, a challenge which, in her estimation, no modern theory (Kantian, utilitarian, or contractualist) was able to meet (Anscombe 1958). The challenge is this. All modern ethical theories use legalistic terminology: they speak of what we are obligated, or duty-bound, to do. In theistic ethical theories, the use of such terms makes sense: God is the moral legislator, and His authority is the source of the normative force behind moral duties and obligations. Modern non-theistic theories, however, want to continue to use these terms in the absence of a moral legislator. So they must find some other way to make sense of them—some other way to ground the authority of morality, in the absence of *an* authority. This is the central challenge which Anscombe has left to modern, non-theistic moral theorists.

Anscombe's challenge: To ground the authority of morality, and thus the legitimacy of our use of legalistic terms such as 'duty' and 'obligation' in moral discourse, without relying on the existence of a moral legislator.

Attempts have of course been made, in the decades following Anscombe's paper, to ground the authority of morality from within the philosophical traditions discussed by Anscombe. I am inclined to think that none has yet succeeded, and thus that Anscombe's challenge stands, to this day, unanswered. It is not my purpose here to discuss and criticize all these attempts; I will limit myself to a brief discussion of my reasons for being dissatisfied with the contemporary rights-based approach to grounding moral duty. Instead, in this chapter and the next, I will sketch an original, comprehensive meta-ethical and normative ethical theory, which I call "Neo-eudaimonism," and argue that it has the resources to meet Anscombe's challenge completely.

2.2 A Starting-Point: Raz's Theory of Practical Authority

Joseph Raz has developed a powerful theory of the conditions under which an individual has a moral duty to obey the directives of an authority. Revising this theory, so that it ceases to be vulnerable to a number of important objections, is the topic of Chap. 15. But for now, we need only sketch the theory in brief, in order to open up a path to answering Anscombe's challenge of grounding moral duties in the absence of an authority.

Raz calls his theory of practical authority "the service conception." It consists of three theses and a definition of 'duty'. The three theses are the Dependence Thesis (DT), the Normal Justification Thesis (NJT), and the Pre-emption Thesis (PT):

> DT: Authoritative directives should be based on the balance of relevant reasons that already independently apply to those subject to the directives (Raz 1986, p. 47).

> NJT: Authoritative directives should make those subject to the authority likely better to comply with the relevant, independently applying reasons by accepting and following the directives as authoritative, rather than by trying to follow the applicable reasons on their own. Demonstrating this is the normal way to justify an exercise of authority (Raz 1986, p. 53).
>
> PT: If DT and NJT are satisfied, then the fact that an authority has issued a directive is a reason to do what is directed which excludes and replaces the relevant, independently applying reasons (Raz 1986, p. 57).

Raz's service conception is an essentially Aristotelian view of practical authority and of the ground of duties to obey the directives of practical authority.[3] As Andrés Rosler has pointed out, Aristotle maintains in the *Politics* that political authority is justified in virtue of the ability of political authorities to serve the interests of those subject to them:

> It is evident, then, that those constitutions that look to the common benefit turn out, according to what is unqualifiedly just, to be correct, whereas those which look only to the benefit of the rulers are mistaken and are deviations from the correct constitutions. (*Politics* III.6, 1279a17–20 translated in Rosler 2005, p. 179)[4]

Duns Scotus, working in the Aristotelian tradition of political philosophy, comes close to formulating the NJT itself:

> Political authority…is just, because anybody can justly submit himself to one person or a community… regarding things in which one can be guided better by him whom one obeys, than by oneself. (*Quaestiones* dist. 14.2. translated in Rosler 2005, p. 178, note 331)

Raz's three theses are to be interpreted in the light of his characterization of duties as *pre-emptive* reasons for action. A pre-emptive reason for action is a special type of *protected* reason. A protected reason, according to Raz, is both (a) a first-order reason to perform (or refrain from performing) some action; and (b) a second-order exclusionary reason; that is, a reason not to act on other first-order reasons which compete with the first-order reason referred to in (a). Authoritative directives create new protected reasons for action. When the first-order component of a protected reason favors performing the act that is supported by the overall balance of pre-existing reasons (i.e., when the authoritative directive that creates the new reason satisfies the Dependence Thesis), we call it *conclusive*. A pre-emptive reason is a conclusive protected reason whose exclusionary component excludes all present competing first-order reasons. This is a duty. When an authoritative directive satisfies the Normal Justification Thesis, the exclusionary reason it creates does exclude all pre-existing reasons against doing what the directive orders (and replaces those pre-existing reasons for doing what the directive orders). Thus, directives that satisfy the DT and the NJT create new pre-emptive reasons—new duties. If there are

[3] In private communication, Raz has told me that he also sees it as the view of authority present in the Talmud.

[4] Rosler also uses Raz's argument for the rationality of obeying authority, which draws on the idea that authoritative directives are exclusionary reasons to disregard one's own first-order reasons to act contrary to the directive, to establish that obedience to authority is consistent with the activity of *phronesis* (Rosler 2005, pp. 90–100).

competing reasons outside the exclusionary scope of the protected reason created by an authoritative directive, those subject to the directive still have a *negative* duty, a duty *not* to disobey the directive in order to accomplish some goal favored only by reasons which are excluded.

According to the service conception, therefore, a legitimate authority is a *de facto* authority whose directives create new pre-emptive reasons for action, and thus new duties, for those subject to the authority (Raz 1986, p. 60). It is in virtue of creating reasons with this pre-emptive structure that the directives of an authority are authoritative, that we are duty-bound to obey them. When one is given an order by one's superior officer to execute a task in a certain way, for example, the order is not just a reason to execute the task in that way, which should be weighed against the reasons for executing it in other ways. That one was so ordered is also a reason to disregard the reasons for executing the task in other ways; the order is an exclusionary reason in addition to being a first-order reason. To fail to recognize it as one is to fail to take seriously the authority of one's superior officer. The authoritative character of the reason thus derives from its pre-emptive structure.

This is why moral duties are identified with pre-emptive reasons for action. When one has a moral duty to do something, one does not simply have a reason to do it which outweighs the reasons against it, or for doing something else. Our moral duties are those reasons which make a special, authoritative claim on us, a claim which we are not free to weigh against all other reasons for not doing our duty. By understanding moral duties as reasons which have an exclusionary aspect, we also capture the notion that when one is under a duty, one lacks the liberty to do something. We are morally at liberty to ϕ so long as the reasons for ϕ-ing are accessible to us—though if those reasons are outweighed, ϕ-ing will not be what we ought to do. Reasons, even if they are outweighed, are accessible so long as they are not excluded. They are excluded when there is sufficient second-order reason not to act on them—that is, when they fall within the scope of an exclusionary reason. Moral duties are pre-emptive reasons because when one is under a moral duty to ϕ, one *both* has a conclusive reason to ϕ (ϕ-ing is what one ought to do, is favored by the balance of reasons) *and* lacks the moral liberty not to ϕ, since all the reasons against ϕ-ing which are currently present are excluded.

Unlike Raz, I maintain the usefulness of distinguishing between obligations and duties. I understand an obligation to be created by a voluntary action of the person who acquires the obligation. The typical example of an obligation is the obligation to keep a promise. One may have reason to do something before one promises to do it—it may even be what one ought to do, all things considered. But when one promises to do it, one does not simply create another reason in its favor. The point of making a promise is to bind oneself so that one is not free to refrain from performing the promised act, even for reasons that would be perfectly legitimate in the absence of the promise. Promising accomplishes this by creating a reason to disregard reasons that compete with the reasons to do what one has promised to do; promises create exclusionary reasons as well as first-order ones. No voluntary action on the part of the duty-bound, on the other hand, is needed to generate a duty. Duties exist when the facts of the situation one finds oneself in constitute a pre-

emptive reason for one to perform an action. I do not claim that all rights are correlative with duties. Some rights are correlative with obligations. I am concerned only with rights that are correlative with duties.

I accept the service conception as basically correct—as I said, we will see in what ways it needs to be revised in Chap. 15. But for now, one might wonder what a theory of moral duty in cases in which there is an authority present has to do with our task of accounting for moral duty when there is no authority figure. The answer to this question is at the very heart of the theory of moral duty I will be articulating. But before we come to it, I should explain why I am dissatisfied with Raz's own rights-based approach to grounding moral duty.

2.3 Raz's Rights-Based Theory of Moral Duty

Raz's central claim is that moral duties are grounded by moral rights. The fact that one individual has a moral right against another is a pre-emptive reason, and thus a moral duty, for the latter to perform the action which will satisfy the right. The focus of Raz's theory, then, is on giving an account of how moral rights are themselves grounded. For Raz, the value of someone's having a right, which derives from the value of satisfying the interest which the right would protect, explains why that person has the right—we say that this value *grounds* the person's right (Raz 1994, p. 45).[5] Not every interest's satisfaction, however, has a value that is sufficient to ground a right of the interest-holder. Raz identifies three features common to interests generally thought fit for protection by rights: the interests are especially important to the interest-holder, they are "relevant to some person or class of persons so that they rather than others are obligated to the right-holder" (Raz 1986, p. 181), and advancing them "serves[s] the common or general good" (Raz 1994, p. 52). He argues that these features make the value of satisfying an interest great enough to ground a right.

Raz takes the first two of these features over from Mill, who in turn followed a venerable tradition in the English common law. Mill's utilitarianism is not essential to his theory of rights; whatever criterion for the rightness of action one uses, one may begin, as Mill does, with the observation that we are not duty-bound to perform the right action in every case (Mill 1863/2002, p. 49). Mill uses the features of importance and relevance to identify the class of right actions which we are under a duty to perform (Mill 1869/1978, p. 79), and then grounds rights on duties (Mill 1863/2002, p. 50).

The first of these two features picks out especially valuable interests, the ones whose satisfaction makes enough of a difference to the interest-holders to make them candidates for protection by rights. The second feature is necessary because

[5] We should, I think, be immediately skeptical of this sort of attempt to explain the existence of moral rights, as it resembles an argument of the form "It would be very good if X existed; therefore, X exists."

every right is a right against some particular person or group, so there must be some person or group that is particularly relevant to the interest's satisfaction against whom a right protecting that interest would be held. Raz's important and original contribution is in observing that the value to the interest-holder of satisfying an interest is often insufficient to justify recognizing a right that the interest be protected. Rights are grounded, he argues, only when satisfying the interest also contributes to the common good, when there is a "harmonious relationship" between individual and public interest (Raz 1994, p. 55). In this case, the combined value to both the interest-holder and the public is sufficient to justify securing the interest coercively, and thus to ground a right. Having derived rights from interests, Raz then derives duties from rights. He claims that rights are "intermediate conclusions in arguments from ultimate values to duties" (Raz 1986, p. 181). Rights are thus a ground of duties, "a reason for judging a person to have a duty, and…reasons for imposing duties on him" (Raz 1986, p. 172). Raz acknowledges that duties may have grounds other than rights, but notes that the duties grounded on rights are significant among duties, in that all duties owed to individuals to advance their interests are grounded on rights of those individuals (Raz 1986, pp. 180, 186).

There are two problems with Raz's account. The first is fairly easy to remedy. We are not always justified in coercing someone who is in a position to do a great good but is unwilling to do it, even if the act in question would advance the common good as well as the good of some particular individual. A fourth feature, which I discuss below, must be added to Raz's list: the negative social impact of securing an interest of that type coercively must not outweigh the good, to both the interest-holder and the public, of the interest's satisfaction.

The second problem, however, is a deep structural one, and it confronts the Razian view (or any view with a similar structure) with a dilemma. The great value of satisfying an interest with the four features just discussed gives the interest-holder and his community a strong reason to secure his interest coercively if necessary. This value will likely outweigh the harm of coercion, including the harm to the person coerced of being deprived of his freedom to determine his own actions. But now we have run afoul of the problem we started with: this is the wrong sort of justification for enforcing a right. In order to escape this horn of the dilemma, the teleological theorist must argue that the interest-relevant in this case is not morally free to refrain from satisfying the interest. He must thus argue that the interest-relevant has an exclusionary reason. Let us try to modify the Razian view in this way, and say that (1) the great value of satisfying an interest with the four features will only ground a right if that interest also grounds an exclusionary reason for the interest-relevant to disregard all reasons not to satisfy the interest. Then the ground of the right would also ground the right sort of justifying reason for enforcing the right, and the constraint would be met. But now we must remember Raz's claims that (2) when duties are correlative with rights the duties are derived from the rights, and that (3) duties are pre-emptive reasons. For the Razian to maintain all three of these claims, he must argue that only once a right against the interest-relevant has been grounded does the overall balance of first-order reasons favor the interest-relevant advancing the interest protected by the right. The right would then be part

of the ground of the pre-emptive reason to advance that interest, and thus be part of the ground of the correlative duty.

In the next chapter, I show that such a position is untenable. An interest cannot ground the needed exclusionary reason without also grounding a first-order reason consistent with the overall balance of reasons. In grounding the exclusionary reason, therefore, the interest actually grounds a duty. This is the second horn of the dilemma: in order to satisfy the justification constraint, the Razian must argue that the great value of satisfying an interest with the four features can only ground a right if that interest first grounds a duty for the interest-relevant to satisfy it. This violates the Razian priority of rights. The grounding of a duty to advance an individual's interest is here part of the explanation of why that individual has a right, rather than the other way around. Raz's theory of moral rights, then, cannot be correct, since it violates the justification constraint; and his theory of moral duty cannot be correct either, since it assumes his theory of rights.

The theory of moral rights I will develop in the next chapter may thus be seen as an extension of Mill's original theory, in which moral duties are given explanatory priority over rights. There is a fundamental difference between my approach to developing a teleological theory and the approach of Raz and other teleological theorists. Interests are a ground of both value and reasons. An interest is a ground of value insofar as a state of the world in which someone's interest is satisfied is, *ceteris paribus*, a good one. An interest is a ground of reasons insofar as the fact that someone has an interest in some act being performed is, *ceteris paribus*, a reason to perform that act. As we have seen, teleological theorists like Raz typically focus on interests as grounds of value, and attempt to ground rights on that value. I focus on interests as a ground of reasons—and of duties specifically—and then ground rights on those duties. We will see that my theory succeeds both in meeting Anscombe's challenge with respect to the ground of moral duties, and in satisfying the justification constraint with respect to the justification of enforcing moral rights.

2.4 Authority, Natural Reasons, and the Principle of Reasons-Isomorphism

The central claim of my theory of natural moral duty—moral duty in the absence of a practical authority—can be summed up by the following *principle of reasons-isomorphism*:

(R-I): Human interests, in the absence of any practical authority, can ground reasons which are similar in structure to the reasons grounded by the issuing of directives by a legitimate practical authority.

In the next chapter I will defend this principle, and answer Anscombe's challenge, by showing how, and under what conditions, interests ground reasons which are isomorphic to those created by practical authorities. Such reasons are authorita-

tive—they are moral duties—in virtue of their pre-emptive structure, just as the reasons created by the directives of legitimate practical authorities are. I will then go on to argue that that these natural moral duties in turn ground moral rights, and do so in a way that satisfies the justification constraint. I devote the remainder of this chapter to making a closer examination of the concept of a moral duty, and to sketching the meta-ethical background of my view of moral reasons.

3 The Structure of Natural Moral Duty

Since I have defined moral duty as a type of reason for action, the reader will doubtless be wondering exactly what I take a reason for action to be, whether a reason for action must move one to act, and why I identify human interests as the ground of reasons. My answer to the last of these questions places me in a particular camp within ethical theory—a broadly Aristotelian one, which is why I call the theory I will go on to develop "Neo-eudaimonism"—and my answers to the first two are likely to be seen as genuinely idiosyncratic. I therefore beg the reader's patience. Rather than beginning with what are, from a certain perspective at least, the most fundamental issues, I will begin by discussing what it means for a reason to be pre-emptive, and what it means for a reason to be exclusionary in particular, and rely on as neutral an understanding of what reasons are as I can—something like the familiar "facts that count in favor of doing something." I do this because what I have to say about the pre-emptive and exclusionary nature of reasons does not rely on my own more fundamental commitments within ethical or meta-ethical theory. My goal is to keep as many readers as I can on board with the view of moral duty I am developing, for as long as possible. But I will not ultimately shirk my responsibility to articulate my commitments on these more basic issues—commitments which form the essential background against which I will develop and argue for my account of how interests ground moral duties. I should warn the reader, however, that my statement of my views on these foundational ethical and meta-ethical topics will only amount to a thumbnail sketch. A full survey of views on the nature of reasons, with criticism of competing theories and full defense of my own position, would take us much too far afield from our current task. I hope, nonetheless, that many will find my background views appealing.

3.1 A Closer Look at Pre-emptive Reasons

For Raz, a single account of pre-emptive reasons suffices, whether those reasons are based on the issuing of an authoritative directive, or the possession of a moral right. But since we have found reason to reject Raz's theory of moral rights, and so his theory of natural moral duties as well, we need a new account of how interests themselves ground pre-emptive reasons (and thus ground moral duties). And we

will see that naturally grounded pre-emptive reasons differ from authority-based ones in a number of ways. These differences, however, leave enough of a structural similarity intact to justify classifying them as pre-emptive reasons and taking them to be authoritative, thus preserving the principle of reasons-isomorphism.

The first difference is that authority-based pre-emptive reasons are what we might call *replacement* reasons. An agent might have any number of reasons for performing some act. But when a legitimate authority directs him to perform it, the fact that he has been so directed replaces his other pre-directive reasons for action. Now it is this fact of being directed, and it alone, which is his reason for acting; and in order to respect the authority's power, it is that reason for which he must act. Raz would want to say the same, within the context of his theory of rights, of the fact that someone comes to have a right against the agent, and the way in which the agent succeeds in showing respect for that right.

Pre-emptive reasons which are grounded by individuals' interests, however, do not replace anything, for the simple reason that there is nothing for them to replace. When an interest has the necessary features—the specifics of which I will discuss at some length in the next chapter—it grounds both a first-order reason and an exclusionary reason. The exclusionary reasons grounded by the kinds of interests capable of grounding duties have wide enough scope to capture some range of commonly encountered competing reasons. The first-order reasons grounded by the kinds of interests capable of grounding duties are normally strong enough to outweigh competing unexcluded reasons. These together constitute a protected reason which is normally conclusive. That is the kind of interest capable of grounding a duty: one which normally grounds a conclusive protected reason. On those occasions when all present competing reasons are excluded, that protected reason counts as a pre-emptive reason, and thus the interest grounds a duty. The first-order component of this pre-emptive reason is just an ordinary, interest-based, first-order reason—precisely the sort of reason which an authoritative directive would replace, if one were to be issued.

From this explanation, we can already see another of the important differences between natural and authority-based pre-emptive reasons. When a legitimate authority issues a directive, he creates a single reason for action which is at once both a first-order reason and an exclusionary reason. This reason is a pre-emptive one so long as all present competing reasons are excluded. But as we will see, this is not how natural pre-emptive reasons arise. Instead, individuals' interests, under the appropriate circumstances, ground both first-order reasons and distinct exclusionary reasons; these reasons, and the arguments which establish that interests do ground them, are separate. A natural pre-emptive reason, then, is a sort of reason-compound: a first-order reason, *plus* an exclusionary reason which protects that first-order reason, both grounded by the same interest.[6]

[6] We can, if we want, be a bit more flexible, and say that a naturally grounded pre-emptive reason consists in a cluster of first-order reasons, and a corresponding cluster of exclusionary reasons which collectively exclude the competing first-order reasons. Each pair of first-order and second-

3.2 The Exclusionary Component of a Pre-emptive Reason

Raz's real breakthrough in the analysis of legitimate authority was to recognize the reasons created by legitimate practical authorities as having an exclusionary dimension, and it is in virtue of having this double-aspect—first-order and second-order—that they are authoritative. Nonetheless, he left many issues regarding the nature of exclusionary reasons unresolved, a number of which are of crucial importance for understanding the character of natural moral duties.

3.2.1 Reasons for Action and Reasons for Belief

Facts about individuals' interests are both reasons for action and reasons for normative belief. That someone I know has an interest in my helping him to achieve a goal is both a reason for me to help him, and a reason for me to believe that helping him is what I ought to do. To be more precise, we should invoke a distinction made by Jonathan Dancy, and say that a reason for action actually stands in two normative relations to an action. The reason *favors* performing the action; and the reason *contributes to the rightness* of performing the action (Dancy 2004). An agent responds to the reason *qua* right-maker by believing that the action is what he ought to do (or rather, by increasing his degree of credence that the action is what he ought to do); and he responds to the reason *qua* favorer by performing the action.

Every fact which is a reason for action is therefore, in virtue of contributing to the rightness of the action, also a reason for belief—a reason to believe (or increase one's credence) that the action is what one ought to do. We should pause to consider the relationships between the concepts of reason for action, reason for belief, and evidence. The evidential relation is one that holds between facts or propositions: fact/proposition 1 is evidence for fact/proposition 2, which is to say that the obtaining of fact 1 (the truth of proposition 1) increases the probability that fact 2 obtains (proposition 2 is true). The reasons-relation is one that holds between facts (propositions) on the one hand, and beliefs or actions on the other: fact R is a reason to believe p, or to do ϕ. And we have seen that when a fact is a reason for action, it stands in two relations to the action: favoring it, and contributing to its rightness.

One fact is evidence for another if and only if the first is a reason to believe (increase one's credence) that the second obtains. The relationship between the concept of evidence and the concept of a reason for action may be more complicated. Every reason for an action is a reason to believe that the action ought to be done; but perhaps not every reason to believe that an action ought to be done is a reason to do it. Stephen Kearns and Daniel Star have recently argued that a fact R is a reason to perform an action ϕ if, and only if, that fact is evidence that one ought to ϕ (Kearns and Star 2008, 2009). Others, such as Mark McBride, disagree (McBride 2013). If

order reasons will be grounded by a single interest, but different interests, and indeed the interests of different agents, may contribute to the compound pre-emptive reason.

an article in a trustworthy periodical states that the people of some country are experiencing famine, that fact (that is, the fact that the article so states) is certainly evidence that I ought to make a donation to a charity that is contributing to the relief effort (rather than, say, a charity whose work focuses on less urgent circumstances), and thus a reason to believe that I ought to do so. Kearns and Starr believe that the fact in question is also a reason for action—that it contributes to the rightness of the act of making the donation, albeit less directly than does the fact reported (the fact that those people are starving). Likewise, the fact that the paper so reports is less direct evidence that one ought to donate than is the fact reported itself (Kearns and Star 2013). McBride disagrees, holding that it is only the fact that those people are starving, not the fact that my newspaper has stated that they are, that is the reason (the right-maker) for that action. I share McBride's view on this point.[7]

A deliberative agent, insofar as he is acting ethically, will only perform an action once he has concluded that it is what he ought to do. If I take a moment to deliberate about whether to help my friend or not, then what I am trying to do is to determine whether I have sufficient reason to conclude that helping her is what I ought to do. Once I have come to this conclusion, I will go ahead and help her. Note that there is no basis for claiming that when I act after deliberating, I act *for* the reason that my action is what I ought to do. On the contrary; I act *for* the reason that favors my action, viz., the fact that my friend is in need of my help. *Qua* deliberative agent, my recognition of this fact moves me to act only once I have concluded that I ought to help him.[8] There is, moreover, no basis for claiming that all action for a reason is

[7] McBride's (and my) position is motivated by an acceptance of John Broome's view that a reason for action *explains* an action's being right (or contributes to such an explanation) in the sense that it makes it the case (at least in part) that the action is right (Broome 2004). In (Kearns and Star 2008), they note that although these two views of reasons seem bound to disagree about what counts as a reason—since normally one fact can be evidence for another without in any sense making the other obtain—it may be an interesting feature of the normative realm that this maxim does not hold there: perhaps what counts as evidence (however indirect) for an action's being right always contributes (however indirectly) to the fact that it is right, and the two theories end up agreeing about what facts count as reasons for what actions. Since, as far as I am concerned, it is non-negotiable that reasons for an action contribute to the rightness of the action, I think that Kearns and Star's view could only be correct if this were so. But I am unconvinced that normative inquiry differs from all others areas of inquiry in this rather extraordinary way.

[8] What precisely does it mean to say that a reason, or the recognition of a reason, "moves one to act"? The orthodox answer in the philosophy of action is that it means the reason (or recognition thereof) causes one's action, in the sense of efficient-causation. This is the mistake at the heart of most contemporary philosophical action theory. The concept of an action belongs to the logical space of reasons, not the natural space of causes. Orthodox philosophy of action mongrelizes—to use Sellars' term—the concept of an action, as it does the concepts of intention and belief, and even the concept of a reason itself. It tries to force these concepts to occupy both spaces simultaneously. To say that one is moved to act by one reason (rather than another) is to say that one has set one's target in acting, that-for-the-sake-of-which one will act, in the way favored by that reason (rather than another), and on the basis of its being favored by that reason (rather than another). Reasons-explanations of action, then, are *forward-looking* versions of *final-causal* explanations. That they are forward-looking is probably why they are so readily confused with efficient-causal explanations. From the perspective of the soon-to-be-acting agent, of course, the action which is to be taken can only be determined in a forward-looking way. Once the action has been performed, its

deliberative—is preceded by deliberation about what one ought to do. I can of course be moved to act by my recognition that my friend needs my help, without having deliberated about whether helping him is what I ought to do. But ethical deliberation does often play an important role in action, and deliberative action will be my focus here. How ethical deliberation fits with ends-deliberation and instrumental deliberation into a single picture of deliberation and action is an important question which I will take up in the next chapter. For now, what we need to examine is just what exclusionary reasons are reasons to do, and what, if anything, they are reasons to believe.

---

performance can be explained either by reference to the reasons which set its target, or by reference to the target itself—that is, either by reasons-explanation or final-causal-explanation. For one of the only attempts that I know of to avoid the mongrelization that plagues so much of the philosophy of action, see (Collins 1987, ch. 6).

I speak in this chapter of being moved to act by a reason, rather of being motivated, since I (idiosyncratically) restrict my use of the term "motivation" to the causal processes which result in voluntary motion. I do this because the contemporary version of the philosophical debate between Humeans and anti-Humeans, which is universally referred to as a debate about motivation and which was discussed in Chap. 4, is concerned with the role of cognitive and affective neurological processes in producing voluntary motion. Of course, there will also be such an efficient-causal explanation for the motion through space and time of the agent's body. There is, however, nothing superfluous about reasons-explanations. They are indispensable to identifying real patterns in the world; our capacity to understand each other, and our ability to predict each other's actions and thus to co-ordinate our lives, would suffer immensely without them. The question of why someone did something, if this is a request for a reasons-explanation, is, like the questions of what someone believes or intends, a fundamentally interpretive one. It is a request to make someone's pattern of behavior—one's own if it is asked of oneself, or someone else's—intelligible. The indispensability of the concept of a reason in this context follows from the similar indispensability of the concepts of belief and action. These concepts all belong to the logical space of reasons; and of the three, the concept of a reason is the most basic. We can only understand what it is to believe or to act if we understand that these are things one does for (what one takes to be) reasons. Much confusion has been born of the fact that when what someone believes is false, we explain (in the sense of making intelligible, *not* in the sense of giving a *causal* explanation) his action by reference to what he believes, and not by reference to his reason for acting—since there was in fact no reason for him so to act (though he believed there to have been). This has led those who labor under the misapprehension that beliefs are causally efficacious states of the brain to the doubly erroneous conclusion that beliefs, rather than reasons, are fundamental to the explanation of action—doubly erroneous, because beliefs are no part of the causal explanation of action (the only sort of explanation they recognize) and are subordinate to reasons in intelligibility-explanations of action. We always act for, and explain our own and others' actions in terms of, apparent reasons, since we have no direct access to the facts. When we explain someone's action in terms of his beliefs, we are signaling the fact that what appeared to him to be the case, and to favor his action, does not appear so to us. And when we explain someone's action in terms of his reasons, we signal that we share recognition of those reasons. Giving a belief-explanation instead of a reasons-explanation distances one from the purported reasons cited in one's explanation, in much the same way that German uses the present subjunctive in indirect speech to distance the speaker from the paraphrased speech. See Chap. 4 notes 9 and 15 *supra*.

3.2.2 Exclusionary Reasons and Wrongness

We have seen that Raz characterizes exclusionary reasons for action as reasons not to act for certain other reasons. But how exactly does a deliberative agent go about responding to such a reason? In order to answer this question, we first have to ask what exclusionary reasons are reasons to believe, and what role they play, as such, in ethical deliberation. Our answer to this prior question invokes one of the insights of Mill into the nature of moral duty, already mentioned above: that we do not have a duty to perform every right action. The right action, the one we ought to perform, is the one favored by the balance of reasons. What is our duty is something more than this; it is what we have a pre-emptive reason to do. I suggest that the key to understanding the nature of exclusionary reasons lies in how we conceive of the *wrongness* of an action. Wrongful actions are not simply all those that are not right, all those favored by reasons which end up outweighed by reasons favoring some other action. They are actions all of whose supporting reasons have been excluded—actions which we are not morally free to perform. First-order reasons may either contribute to or detract from the rightness of an action; but only exclusionary reasons contribute to the wrongness of an action. They stand in a wrong-making relation to actions, and are thus reasons that support the belief that the actions favored by the reasons they exclude are wrong.

The chief advantage, as I see it, to understanding the wrongness of actions in this way is that the alternative requires a fairly inhumane conception of morality. There are many situations in which what we ought to do, what we have most reason to do, is something that would require a great deal of self-sacrifice. I find myself unable to accept a view of morality according to which any time a person fails to do precisely what he ought, all things considered, he has committed a wrong and deserves, if not punishment or blame from fellow men, then at least to be harried by feelings of shame or guilt.

It is no surprise then that this understanding of wrongness also allows us to recognize and give a satisfying account of the existence of supererogatory actions. It would take us too far afield to go into this topic in any depth here.[9] But we can briefly note that understanding wrongness in this way opens up the possibility of an account of supererogation which satisfies a couple of elusive desiderata. First, it becomes clear why we should have a *strong* account of supererogation—one according to which a failure to perform a supererogatory act is not a wrong, as opposed to being an excused wrong. It allows us to define supererogatory acts as self-sacrificing acts, performed to advance the interests of others, which are favored by the balance of reasons but which are not duties. Since supererogatory acts are not duties, the circumstances that make them possible are ones in which there are non-excluded reasons for performing other, less demanding actions. And so we need no additional theoretical apparatus, no special permission to forgo the supererogatory

[9] I do so in my "Reason, Virtue and Supererogation: The Unfinished Project of 'Saints and Heroes'" (Unpublished MS).

act even though it is right, in order to see why such failures are not wrongs.[10] They are not wrongs because we can fail to perform these acts without thereby acting for an excluded reason. Second, it allows us to maintain that supererogatory actions are right, as opposed to merely being actions of great moral value which are nonetheless no more favored by the balance of reasons than other, less taxing options. Thus we avoid turning the supererogatory into something quixotic—an act of self-sacrifice which, however beneficial, is not ultimately what one ought to do, any more than some other less costly act.[11] We will then have no trouble making sense of the fact that supererogatory actions are particularly deserving of moral praise and encouragement, since we can acknowledge them as favored by the balance of reasons over their less-demanding alternatives.

So exclusionary reasons are reasons to believe that the actions whose supporting reasons they exclude are wrong. But it is of course possible for some reasons favoring an action to be excluded, while others are not. As I explain in the next chapter, interests can ground exclusionary reasons of varying scope, depending on what features they have. And I hold to the view that wrongness is a property of actions, not a property of agents or motives. If some of the reasons favoring an action are left unexcluded, then the agent who performs it cannot be said to do the wrong thing. So an exclusionary reason that excludes some of the reasons for performing that action is not a conclusive reason to believe the action is wrong; it will only be so if it excludes them all. It is, however, a conclusive reason for the agent to believe that he will be acting *badly* if he performs the action *for* one or more of the reasons that are excluded. This is a way of cashing out the familiar thought that it is possible to "do the right thing for the wrong reason." In fact, we should take this thought quite seriously, and allow that an individual's interest may ground second-order reasons which exclude first-order reasons that favor the *same* action as do the first-order reasons grounded by that very interest. If I am the only one able to do a friend a crucially important favor and I do so, but *solely* for the reason that my act will mean that she is subsequently in my debt, then I have acted selfishly—and so acted badly—even though I have done the right thing. My friend's interest excludes that reason for helping her, even though helping her is what I ought to do. The argument for this sort of claim—the explanation of why, and under what conditions, exclusionary reasons are grounded by interests, and how to determine what other reasons they exclude—will come in the next chapter. For now, our concern is simply to get a firm grasp of the precise nature of this type of reason. Exclusionary reasons not

[10] Raz attempts to explain the supererogatory by positing just such a special permission (Raz 1975). But as Dancy points out, it is hard to see what fact could ground such a permission. He suggests that it could only be something like "this is very demanding," and points out that this fails to recognize that the supererogatory is a matter of degree (Dancy 1993, pp. 127–143). There is no threshold of demandingness, below which there are only duties and above which the best acts are all supererogatory. Supererogatory action is possible at any level of self-sacrifice—a point Urmson makes when he characterizes such acts as "every case of 'going the second mile'" (Urmson 1958, p. 205).

[11] This is the problem that plagues Dancy's view of supererogation.

only contribute to the wrongness of actions; they can also stand in a bad-making relation to the motives of agents.

3.2.3 Exclusionary Reasons and Ethical Deliberation

Now that we understand exclusionary reasons as reasons for belief, we can inquire into the role they play in ethical deliberation, and what it means for a deliberative agent to respond to them as reasons for action. Recall that it is the pre-emptive structure of authority-based reasons—the fact that they are both reasons to do something, and reasons to exclude the reasons against doing it—that makes them authoritative. The reason created by a legitimate authority's directive is not simply one more reason to be weighed against whatever reasons there may be for acting otherwise. This point translates to the case of naturally grounded pre-emptive reasons. The deliberating agent is interested in determining both what he ought to do, and whether he has a duty to do it. So he is interested in what interests are present, whether any of those interests ground exclusionary reasons, and where the balance of unexcluded first-order reasons lies. Let us assume that there is at least one unexcluded reason. In determining what he ought to do, then, the agent is not to weigh the reasons that are excluded in the balance against the unexcluded reasons. This is not to say, as if through some normative magic, that they cease to be present, or to have the weight that they have. But the agent passes over them in his deliberations, because the fact that they are excluded makes attending to them in ethical deliberation inappropriate. The question of what he ought to do overall becomes the question of where the balance of non-excluded reasons lies, and to make this determination becomes the focus of his deliberation. If the non-excluded reasons all favor the same action, then the agent has a natural moral duty to perform that action: all present opposing reasons are excluded, and the balance of non-excluded reasons is clear. But it may be that although many reasons against performing some action are excluded, and the balance of non-excluded reasons clearly favors it, still some reasons against remain in play. These are the cases in which performing the better action is supererogatory. This is the scenario in which the agent has a conclusive protected reason, but not a duty.

This account seems to leave us with the following, perhaps uncomfortable, possibility: that in a given situation, the balance of *all* first-order reasons, including those that are excluded, may lie in favor of one action, while the balance of unexcluded reasons lies in favor of another. But as I will be at pains to show in the next chapter, the sorts of first-order reasons grounded by interests that also ground exclusionary reasons are very strong. They are normally strong enough to outweigh even those competing reasons which fall outside the exclusionary scope of their corresponding second-order reasons—if there are any—which tend to be stronger than the excluded ones. And the interests which ground second-order reasons with wider exclusionary scopes also ground stronger first-order reasons. So even if the reasons that get excluded were to be counted, they would never be weighty enough to tip the balance. This uncomfortable situation cannot arise.

Therefore, just as a legitimate authoritative directive commands that the action favored by the balance of pre-existing reasons be done, a naturally grounded moral duty is in fact a reason to do what the balance of all reasons favors. That it is right to do one's duty, however, is not explained by this, but by the fact that one's duty is what is supported by the balance of unexcluded reasons. As far as locating the balance of first-order reasons is concerned, it makes no difference whether excluded reasons are counted or not. Nonetheless, the correct way for an agent to determine what is right is by finding the balance of unexcluded reasons. In neglecting excluded reasons in his deliberations about what action is right, the agent is responding—correctly—to the second-order reasons that exclude them. This means that he neglects them because counting them is inappropriate in the circumstances, not because they make a difference to the balance of reasons one way or another.

Let us now suppose that an agent has determined through deliberation that he is in a scenario in which he does not have a duty to do what he ought to do, but some reasons against doing it are excluded. What does it mean for the agent to respond to these exclusionary reasons as reasons for action? First let us suppose that there is some action the agent could perform, all of whose supporting reasons are excluded. The exclusionary reasons make that action wrong; we might say that the agent here has a negative duty, a duty not to perform this action, even though, *ex hypothesi*, there is no action he has a duty to perform. For the agent to respond to these exclusionary reasons as reasons for action is for him to shape his own choice situation, at the time when he is deciding among courses of action, in the way that they recommend—to disregard, or to ignore, the excluded reasons and the openness of the course of action they support. He responds to them by adopting the pretense that the first-order reasons they exclude are not there, and the action those excluded reasons support is not a possibility for him.

Suppose now that there is some other action, some of whose supporting reasons are excluded and some not. The exclusionary reasons make it the case that to perform that action for one of the excluded reasons—to have the consideration or recognition of one of those reasons as what moves the agent to action—would be to act badly. This case is similar to the first: the agent responds to the exclusionary reasons by ignoring what they exclude, by focusing his attention on unexcluded reasons and trying to make those motivationally salient for him, if they are not already.

3.3 Reasons for Action

I have defined moral duties as pre-emptive reasons for action. But such a definition is of little help without making clear what is meant by a reason for action generally. Thus far I have taken the familiar view that reasons for action are facts which count in favor of performing some action. This familiar view leaves us with a number of deep questions. I will now raise three of them, and provide a succinct statement of the answers to which I am committed.

3.3.1 What Are Facts?

By "the facts" I simply mean all the true propositions.[12] This view of facts as propositions does a most straightforward and economical job of handling a host of problems concerning the nature of facts, especially that of accounting for the existence of negative facts, such as its not being the case that my dog is lying on the rug. But as soon as we adopt a view of facts as propositions, we must have something to say about what propositions are. On this point, I am inclined to accept Scott Soames' recently developed "Cognitive-Realist" account of propositions, according to which a proposition is a cognitive event-type, where the cognitive event in question is one in which something is predicated of something else, as occurs whenever one has a perceptual experience of *a* as *F*, or adopts a belief that *a* is *F*, or understands an utterance "*a* is *F*" (Soames 2010, ch. 6). Soames' view seems to me to go quite a long way toward demystifying the nature of propositions, and the reader may take me to be committed to it.

With this view of facts-as-true-propositions in place, we can spell out more fully what it means to recognize a reason, to recognize a fact *as* a reason to act, and to act for a reason. Suppose the reason in question is the fact that my friend is in need of my help. My recognizing this reason then refers to the specific cognitive events of my perceiving my friend as in need of my help and judging that my friend is in need of my help. My recognizing this fact as a reason to act refers to the cognitive event of judging that the fact that my friend is in need of my help is a reason to help her. This second judgment is the one that plays a role in my ethical deliberations, should I engage in any. For me to act *for* the reason that my friend needs my help is for me to be moved to act in light of my recognition of that fact, whether or not I have recognized that fact *as* a reason for action, and whether or not I have engaged in ethical deliberation concluding in the judgment that I ought to help my friend.

Since I take pieces of evidence to be facts, I identify them with propositions as well. So when the ends-deliberator of the early chapters judges one end more choiceworthy than another, and does so on the basis of pieces of evidence such as the fact that the one is achievable while the other is not, or the fact that that pursuing the one is accompanied by a positive emotional experience but not the other, he is adopting beliefs in the truth of propositions of the form "this end is achievable," etc., and updating his probability judgments on the value of the ends in question accordingly. This much of the all-proposition ontology of the background decision theory I endorse fully, and not merely as a theoretical convenience.

3.3.2 What Is It for a Fact to *Count in Favor of* an Action?

The familiar characterization of reasons for action is hardly an illuminating one, even if we assume that we have a good grip on what facts are. We do not really *explain* what reasons are when we say that they are facts which "count in favor" of

[12] The view that reasons for action are true propositions seems to be Aristotle's own (*NE* VI.2 39b13; Reeve 2013, pp. 119–120).

something, however true this may be. To say that fact *X* favors performing some action ϕ is really just another way of saying that fact *X* is a reason to ϕ. Dancy, who rightly points out that meta-normative facts of this sort ("the *fact* that the fact that *X* is a reason to ϕ") are the most basic normative facts, takes them to be non-natural states of affairs (Dancy 2005a, p. 137). But on my view, to say that this meta-normative fact obtains is just to say that the proposition "the fact that *X* is a reason to ϕ" is true. When we ask, then, what it means for a fact to favor an action, what we really want to know is what makes such a proposition true, and under what conditions is it true. I briefly postpone giving my answers to these questions, which I will provide in the course of sketching my own background meta-ethical views. Our next question concerns the relationship between reasons for action and motivation to act.

3.3.3 What Is It for a Fact to Count *for an Agent* in Favor of an Action?

I do not assume that for A to have a reason to ϕ, A must actually be moved to ϕ. A fact that favors ϕ-ing can be a reason for A to ϕ so long as (1) it is possible for A to be moved to ϕ by his recognition of that fact, given that A cannot transcend the limits of normal human psychology; and (2) A has, or is capable of acquiring, the physical resources necessary for ϕ-ing. The sense of "possible" in (1) is important. We need not assume that in order for a fact to be a reason for an agent to act, it must be capable of moving that agent to act *holding fixed* his actual, current preferences and goals. We need only assume that it is capable of motivating the agent, given some set of preferences and goals which the agent is psychologically capable of acquiring. We can thus maintain, with Bernard Williams, that every reason for action is capable of being invoked in an explanation of an agent's action, without having to deny (implausibly) that a fact is a reason for action for an agent if his recognition of that fact fails actually to move him to act (Williams 1981). We can simply say that it must be possible, in the way just described, for any reason for action to be aligned with some possible element of the agent's subjective motivational set. The connection between reasons and action-explanation which Williams rightly draws our attention to does not require any more than this.

By understanding what it means for a reason to be capable of playing a role in action-explanation in the way proposed here, we steer between the Scylla of extreme internalism about reasons for action, and the Charybdis of extreme externalism. And so we can maintain, for instance, that a healthy agent who has made correct judgments about what ends to pursue is rationally required to exercise self-control in the face of temptation, and that that rational requirement is a normative requirement. It is no objection to point out that there may be situations in which no agent, however healthy, however able he may be to manipulate the structure of his choice situation, and however well he has developed his powers of self-control, would be able to resist some temptation. If there are such situations, it is not the case that the agent ought to resist temptation in them, as the agent lacks reasons to do so. And if he lacks reasons to do so, then he is not correct in judging that he ought to.

4 Neo-eudaimonism Part I: Meta-ethical Background

I have indicated that in order to meet Anscombe's challenge, we will need a new meta-ethical and normative ethical theory. I now begin the task of sketching this theory, which I will complete in the next chapter.

4.1 Interests as the Ground of Reasons for Action

A person's interests are aspects of his well-being. An agent has interests in the formation, pursuit, and achievement of valuable goals and relationships (or more generally, of functionings). We may identify valuable goals as those which the individual could adopt at the conclusion of a well-executed course of ends-deliberation. One's interests, therefore, are the valuable potential ends which one actually adopts through good ends-deliberation, or which one would adopt if one were to deliberate about ends well; but we include as ends not only achieving valuable goals and relationships, but also deciding which ones to pursue and pursuing them, and in particular, doing all these freely and autonomously. The extent of one's well-being is the extent to which one freely and autonomously decides on, pursues and achieves valuable goals and relationships—the extent to which one's interests are fulfilled. For an individual's interest to be fulfilled, for a contribution to be made to her well-being, is itself good; and the fact that an action will advance an individual's interest is the kind of fact that can make it the case that that action is a good thing to do. Interests in general, the potential ends which could be adopted by some agent at the conclusion of a well-executed course of ends-deliberation, are thus a ground of value—they are bearers of value and objects of proper valuing, though I do not claim that human well-being *exclusively* is valuable.[13] This is a broadly eudaimonistic conception of well-being. A life of developing one's capacities and using them to pursue and achieve goals which are freely chosen on the basis of excellent ends-deliberation is a flourishing life, a life of *eudaimonia*, as this notion is understood within the Aristotelian tradition. Advancing a person's interest makes a constitutive contribution to her well-being: it moves her closer to forming, or being able to pursue, or achieving (at least) some one of his valuable goals.

Interests are also the ground of reasons for action. To say, on some particular occasion, that some person's interest is the ground of a reason for action is to say that, on that occasion, the fact that an action would advance that person's interest is a reason to perform that action. This, for the eudaimonist, is where our reasons for action come from—from facts about how our own interests, and the interests of others, are advanced. Since we are taking moral duties to be reasons of a certain type,

[13] I am thus in agreement with the eudaimonism of the environmental philosopher John O'Neill (O'Neill 2002, ch. 1).

the theory sees moral duties as existing in virtue of facts about human interests, and thus facts about aspects of human flourishing.

In identifying human interests as the common ground of both reasons and value, I reject so-called "buck-passing" views of the right and the good, which take one of the two to be dependent on, or derived from, the other. Rather, like Dancy, I see reasons and values as coeval—distinct but proceeding from the same ground (Dancy 2005b). In fact, I find reflection on this issue of buck-passing to be the clearest and most reliable way of getting at the essential differences between rival theories in normative ethics. The essential distinction between deontology and consequentialism is that the former identifies the rightness of actions as the ground of the goodness of states of affairs, while the latter identifies the goodness of states of affairs as the ground of the rightness of actions (different deontologists may of course disagree about the criteria for right action, but they will all agree on this point *qua* deontologists). Recall that this is why I was able to argue in Chap. 3 that a deontological theory can be given an expected value representation just as well as a consequentialist one can: the formal apparatus of decision theory is neutral about this issue of buck-passing. Both deontology and consequentialism are buck-passing views. There are three types of moral theory which, by contrast, refuse to pass the buck. One is Ethical Intuitionism, the theory to which Dancy subscribes, which has its origins in turn-of-the-century Oxford, with H.A. Prichard's development of a normative theory on the foundation of John Cook Wilson's epistemology (Prichard 2002; Wilson 1926). Intuitionism identifies non-natural facts as the ground of both the rightness of actions and the goodness of states of affairs (individual intuitionists may disagree about whether goodness and rightness are grounded by the same sort of facts or different sorts, and about the content of these facts). Another theory that does not pass the buck between goodness and rightness is virtue ethics. For the virtue ethicist, the possession and expression of the virtues becomes the ground of goodness, rightness, and human well-being. Finally, there is my own camp of eudaimonism (and how this theory differs from virtue ethics is a topic I address below). For the eudaimonist, human well-being is the ground of goodness and rightness, as well as of what counts as a virtue and the expression of a virtue.

4.2 Neo-eudaimonism as Aristotelian Pragmatism

4.2.1 *Phronesis, Eudaimonia,* and the Normative Order

Neo-eudaimonism begins with two distinctions. The first is between a conception of practical reason, a conception of the good life, and a conception of the web of reasons for action that apply to agents. The second distinction is between the initial state of each of these—our initial understanding of the canons of practical reasoning, of what it means to lead a good life, and of what reasons for action there are—and the final, in a sense of "final" still to be explained, state of each: what we may call, respectively, *phronesis*, *eudaimonia*, and the Normative Order. The substance

of the view is in the story it tells about what it means to progress from the initial states of these conceptions, to their final states.

Our initial states of these conceptions are not devoid of content, though the content they come with is quite general. Our initial conception of practical reason includes the principles of logical and Bayesian inference. For the neo-eudaimonist, *qua* Aristotelian, our initial conception of the good life includes those very general constituent final ends which are *indispensable* for a human being. These are the ends which must be components of any conception of the good life which we, as human beings, could possibly recognize and endorse as a conception of the good life, given the kind of creatures that we are—physically, psychologically, and socially.[14] Recall the list of types of evidence which bear on the choiceworthiness of any potential end. In the context of Chap. 4, that list was meant to be ecumenical. It was meant to include all the varieties of evidence which would be endorsed by the proponents of each major ethical theory. A proponent of any given ethical theory is free to delete certain items off the list, and to interpret those that remain from the perspective of his preferred theory.

The neo-eudaimonist accepts every item on the list. He does so, moreover, because he believes that those forms of evidence mirror the indispensable constituent final ends. This is the distinctively neo-eudaimonistic explanation for why each of those types of evidence belongs on that list. Any good human life is a life in which one's capacities are developed and exercised; a life of achievement; a life of emotional satisfaction; a life of coherent pursuits; a life of exchanging views with others; a life of benefitting others; and a life lived in response to reasons.[15] It is so because of the physical, psychological, social and normative creatures that human beings actually are. And thus our initial conception of our final end is, according to the neo-eudaimonist, something a bit more contentful than the thin end of leading a worthwhile life, whatever that might be. We begin, unavoidably, with interests in leading a life that possesses each of those features; and so we begin with a conception of the good life that has those constituents.[16] I have already stated that the neo-

[14] Another even more Aristotelian way of putting this is to say that while the ability to reason comes with the development of our second nature, what can count as a reason for us depends on our first nature. Roger Teichmann makes this point (Teichmann 2011, pp. xii–xiii).

[15] Every item on the list should be familiar from Aristotle's *Ethics*. The development and successful exercise of one's capacities in accordance with right reason is the core of Aristotle's conception of *eudaimonia*. The importance of benefitting others, exchanging views, and having positive emotional experiences are brought out in the discussions of friendship and pleasure. The importance of achieving coherence in one's views and ends is revealed in the method of ethics, which seeks to resolve dialectical puzzles about the good life while preserving as many credible common beliefs as possible.

[16] From an Aristotelian perspective, the question of what are the indispensable general constituents of any good human life is one that belongs to the "science of man"—the intersection of all the natural and social sciences that study human beings and aspects of human life—and the one who knows the answer possesses practical *nous*—a component of theoretical wisdom. See the discussion of Aristotle's theory of practical reasoning in Chap. 4 *supra*. These indispensable general ends are not themselves objects of deliberation—we discover them, rather than choose them, as Aristotle thought we discovered what it means to flourish in full. Possession of this practical understanding

eudaimonist takes human interests to be the ground of reasons for action. So given that we all start out with a set of very general interests, we also start out enmeshed in a web of very general normative requirements. We all have reasons to pursue those very general interests in some way, and to aid others in their pursuit (in some way) of those same very general interests. For a pluralist Aristotelianism, the idea of a life in which these indispensable ends are achieved, as the general form of any good human life, is the *arche*, the starting point, of ethics, insofar as human interests in achieving these ends are the ground of reasons for action.

The process of moving beyond our initial conceptions of practical reason, the good life, and the web of reasons, is one of construction—though as I will explain in Sect. 4.3 below, I am *not* a Constructivist (with a capital "C"), as this term is typically understood in meta-ethics. Our conception of practical reason is immediately enriched by our recognition that the interests of others ground categorical reasons for us. An essential aspect of good practical reasoning is taking one's normative situation into account in determining what ends to adopt and pursue. The capable ends-deliberator selects ends whose pursuit is normally consistent with respect for the balance of categorical reasons that apply to him, as I discussed in Chap. 4. But to do this, he must have a good grasp of what those reasons are and how he should weight them. So in order to deliberate about ends well, one must be skilled not only in identifying the categorical reasons grounded by others' interests, but also in determining how they should be weighted, and what their exclusionary scope is. So the activity of practical reasoning is not limited to reasoning about ends and means. It includes ethical deliberation about the weight and exclusionary scope of reasons, an activity which is necessary for good ends-deliberation. The *phronimos* will be the agent who embodies excellence in ethical deliberation, along with deliberation of the other types. I discuss the relationship between these types of deliberation at the end of the next chapter. Ethical deliberation, moreover, is even more of a social, dialogical process than ends-deliberation. Ends-deliberation does involve testimonial evidence, as discussed in Chap. 4; but it does not have as its goal a convergence with others on what it means to lead a good life, beyond mutual respect for the different specific conceptions of the good life reached by different agents who deliberate well about ends from their own positions. Ethical deliberation, on the other hand, does aim at convergence on the content of the Normative Order, and this convergence requires dialogue about our interests and the reasoning that led us to them, and in which conclusions about the weight and scope of the reasons our interests ground for each other are worked out collectively and cooperatively. Again, this is a point I shall return to in the next chapter.

Each agent begins to use his practical reason to determine how he will fill in his conception of a good life, and which specific ends to adopt. He does so by engaging in ends-deliberation on the basis of the categorical reasons he has recognized, along with what he learns about his own talents, emotional responses, social interactions,

---

is what the possibility of excellent ends-deliberation, and thus *phronesis*, rests on. In Aristotle's theory of practical reasoning (and, we now see, in my own), practical wisdom has access to, and makes use of, the truths of theoretical wisdom.

etc.—his "experiments in living," to use Mill's phrase. This is why the neo-eudaimonistic conception of well-being is a pluralistic one. There is no one set of human interests, and thus no one monolithic archetype of well-being, of the flourishing human life. There are as many versions of the good life as there are different courses of well-executed ends-deliberation, carried out by differently situated agents. Our final conception of *eudaimonia*, then will not consist of a conception of a single good life, but rather of the set of all possible complete conceptions of a good life.[17] And that is the neo-eudaimonist's particular reason for asserting (as I did in the Chap. 5) that a good human life must be a life of liberty. Many of the elements of a good life for a particular agent must be chosen by that agent, and are only constitutive of a good life *for that agent* in virtue of being chosen *by him* from among the total set of ends he could have adopted on the basis of deliberation. The very possibility of constructing *one's own* conception of a good life, and thus of leading a life of one's own that carries out that conception, requires the liberty to deliberate about, choose, and pursue particular ends.

Each agent then begins to actually fill in his conception of a good life on the basis of the conclusions reached in those deliberations. The adopting of new ends on the basis of such soundly executed deliberation generates new interests, and thus new non-categorical reasons for action for the one who adopts them. And the new interests generated by the practical reasoning of others in one's community then ground new categorical reasons for action for each agent—reasons whose weight and scope must be worked out through ethical dialogue. These new reasons for action in turn further enrich the standards for good practical reasoning. The agent must take them into account in determining what further, or more specific, ends to adopt and pursue, and what actions to take in order to pursue them. And on the process goes, with the expanding normative web adding to the stock of facts that must be incorporated in practical reasoning, practical reason constructing a fuller conception of the good life on the basis of both the expanding normative web and the personal discoveries made through experiments in living, and the adoption of fuller conceptions of the good life generating new interests which ground new elements in the normative web.

In claiming that the interests of others ground reasons for action which one must take into account in one's own practical reasoning about what ends to adopt and how to pursue them, I have obviously skirted over some very difficult questions. What any agent must determine is not simply that the interests of another ground reasons for action for him, but what weight to give those reasons, and what exclusionary scope, if any, to recognize them as having. How is the agent to do this? What canons of ethical deliberation is he to use, and how is he to apply them in his ethical dialogue with other agents, in determining the weight and scope of the very normative reasons which he must then in turn incorporate into his own deliberation

[17] We can thus see the goal of articulating a complete conception of *eudaimonia* as a communal correlate of the individual's aspirational goal of completing his own conception of a good life for him, as discussed in Chap. 4. As I discuss below, a complete conception of *eudaimonia* is an aspirational goal in an even more radical sense.

about his own ends? The theory of how interests ground moral duties and moral rights, and of the structure of moral reasoning and dialogue, which I will go on to develop in the next chapter contains my answer to these questions. The agent makes these determinations by reflecting on certain features of the interests that compete for his attention—those features mentioned above, such as importance to the interest-holder, which I will define more precisely and discuss in greater detail below. One specific feature which will play an important role in determining the strength and scope of the reasons grounded by an interest is its centrality to the agent's plan of life—which is simply a function of how highly ranked the pursuit of satisfaction of that interest is (or would be if the agent were to deliberate well about ends). So just as excellent ends-deliberation requires excellent ethical deliberation—just as determining the values of one's potential ends requires determining the structure of the web of reasons for action that apply to one—excellent ethical deliberation requires an awareness, achieved through dialogue, of the conclusions others have reached in their ends-deliberations, since the weight and scope of the reasons that apply to one are partly determined by the value attributed to the interests that ground them by those who have deliberated well about their ends.[18]

My theory of moral duty will remain at a certain level of generality, leaving many questions of finer detail undecided. The project of determining the strength and shape of the normative force exerted on us by the interests of others is, for all of us individually and collectively, a work in progress, *as indeed it must be*. I will now say a few words about why I believe this to be, and in so doing, will answer a few other fundamental questions which have been looming over the discussion so far.

4.2.2 Normative Pragmatism

In defining reasons as facts and facts as true propositions, I raised the question of what makes normative propositions true. The answer I give to this question is essentially the same as the one I would give to the question of what makes the propositions of scientific theories true. It is the answer of a Peircean pragmatist:

> All the followers of the method of science are animated by a cheerful hope that the process of investigation, if only pushed far enough, will give one certain solution to each question to which they apply it…Different minds may set out with the most antagonistic views, but

[18] Note that this recognition of the mutual dependence of ethical deliberation and ends-deliberation does not violate the neo-eudaimonist's commitment to rejecting buck-passing accounts of rightness and goodness. The achievement of an end is not valuable in virtue of the fact that it results from a right action; rather, the achievement of an end which does result from right action is valuable, at least in part, in virtue of the fact that that achievement satisfies the indispensable interest in leading a life in accordance with reasons. That the action which led to the achievement was a right action is evidence of this being the case. And although the value of an interest may, as will be discussed in greater detail in the next chapter, intensify the strength of the reason for action grounded by that interest, the action that advances it is right in virtue of being favored by the balance of reasons, not in virtue of being the action that advances the interest of the greatest value.

> the progress of investigations carries them by a force outside themselves to one and the same conclusion. This activity of thought by which we are carried, not where we wish but to a foreordained goal, is like the operation of destiny. No modification of the point of view taken, no selection of other facts for study, no natural bent of mind even, can enable a man to escape the predestinate opinion. This great hope is embodied in the conception of truth and reality. The opinion which is fated to be ultimately agreed to by all who investigate is what we mean by the truth, and the object represented in this opinion is the real. That is the way I would explain reality. (Peirce 1894, 407)

The true normative propositions, then, are those around which the community of normative inquirers would form a stable consensus at the hypothetical end, the ideal limit, of an infinite process of normative inquiry—the process of collectively determining what evidence, what reasons, there are that one action or another is the one that ought to be done by an agent in a given position in a given set of circumstances, and of working out the scope and weight of those reasons. They are true, moreover, *in virtue of being so*.[19]

---

[19] A pragmatic conception of truth as I understand it—whether for normative or scientific propositions—is *not* one that understands truth as transcending any conceivable context of justification. This fact is often cited as in itself a defect in the pragmatic theory of truth, and as solid ground for rejecting it. But from a pragmatic perspective, it is precisely this dogged insistence that any adequate concept of truth must be justification-transcendent that is mistaken; it is an article of faith which we ought to struggle to free ourselves from. The essentials of this pragmatic view of truth were shared by a number of twentieth century thinkers who were equally influenced by both Kant and modern science, most notably the Marburg neo-Kantians (especially Ernst Cassirer), the great French mathematician and scientist Henri Poincaré, Hilary Putnam during his "internal realist" phase (roughly 1981–1990), and the Critical Theorist Karl Otto-Apel. In his later work, Putnam turns back to thinking about truth in a justification-transcendent way, as does Habermas, who had endorsed a non-transcendent, though not really pragmatic, theory of truth in much of his work. I should be careful to note that Peirce's own views on truth, reality, and the relationship between the two changed considerably over the course of his life. For an excellent historical study of this development, see (Hookway 2000). My own view takes not only the concept of truth, but also the concept of reality itself—both the parts of reality described by scientific theory and the parts described by normative theory, which I take to be complementary and to form a larger whole—to be a regulative ideal, and this is a combination of views which Peirce may never have held simultaneously. To say that our concept of reality is a regulative ideal is to say that the only intelligible and useful concept of reality we have is a concept of reality-as-we-experience-it, and reality-as-we-experience-it is simply that which would be described by the scientific and normative theories endorsed at the ideal limit of the infinite process of scientific and normative inquiry. Any conception of reality which reaches beyond this transgresses the bounds of reason and experience by trying to give content to the purely negative regulative idea of reality-as-it-is-anyway, which can only be the idea of a *Ding-an-sich* (or, if structuralism is the philosophy of science we ought to adopt, a *Struktur-an-sich*). I therefore find not only the metaphysical realist's view of the world, but also Bernard Williams' idea of the absolute conception of the world (whose differentiation from metaphysical realism is debatable), to be neither necessary nor intelligible nor useful. The ambition of both scientific and normative inquiry, which I take to be joint endeavors, does not require the goal of arriving at an absolute conception, but is to be understood in terms of a hopeful search for convergence on a single conception of the world which is adequate to the full range our experience. If this is not, in the end, an interpretation of his writings that Peirce himself would endorse, then so be it. It is certainly the view of Poincaré and Cassirer.

There are a number of bad objections to a pragmatic theory of scientific truth—even one which understands truth as a merely regulative ideal, and so takes convergence at the limit of inquiry as

Some elaboration on this idea of truth-as-consensus-at-the-limit-of-inquiry is needed in order to see how it can be made to apply to normative ethical propositions, in addition to propositions of science.[20] The first point to note is that scientific truth does have an important bearing on normative ethical truth. The relationship between the two, which must be recognized by any intellectually responsible ethical theory, is summarized nicely in what Owen Flanagan calls the Principle of Minimal Psychological Realism (and which I would prefer to call the Principle of Minimal Biological, Psychological and Social Realism):

---

something to be hoped for, not something that would necessarily be achieved—which are unfortunately common. The first considers the possibility that we will reach a point when the evidence required to decide between rival scientific theories becomes permanently inaccessible to us. For example, we may reach a point when deciding between rival theories in physics would require conducting experiments at energy levels which are impossible for us to reach. The second, related, objection considers the possibility that we will reach a point where our cognitive capacities, given the type of evolved, biological creatures we are, are simply incapable of formulating a successor theory which would reconcile the inconsistencies between our current theories and account for all the evidence they fail to make sense of. In both of these cases, we will have run up against an insurmountable barrier—technological in the first, cognitive in the second. The point of the objections is that both of these cases are consistent with there being a possible theory which we would converge on, if only we could formulate it/get at the evidence for it. But even given infinite time for inquiry, we never would converge on it. The problem with both of these cases, however, is that they depict scenarios in which we are prevented by an insurmountable barrier from ever reaching the limit of inquiry, even given infinite time for inquiry. And so they do nothing to undermine the conception of truth-as-convergence-at-the-limit-of-inquiry as a regulative ideal. A third objection claims that a competent inquirer could go so wrong at some stage of his inquiries that he is never able to find his way to the conclusion that those inquiries would have converged on, had he not gone so far off course. This objection simply misunderstands the notion of Bayesian convergence. Given a partition of mutually exclusive hypotheses, the accumulation of evidence will eventually send the probability of one hypothesis to 1 and all the others to 0, regardless of the inquirer's priors. This will only fail to be the case if (a) a genuine partition of hypotheses is impossible, because one or more cannot be formulated given the cognitive capacities of the inquirer; (b) at some point the remaining necessary evidence becomes permanently inaccessible; or (c) reality as the inquirer experiences it is itself fragmented and inconsistent. In the case of this third possibility, the pragmatist must concede that there is no complete, self-consistent truth. I return to this point briefly below. Finally, there is the objection that even if we would converge on a single theory at the limit of inquiry, that theory might still be *wrong*. This objection is simply a flat rejection of any non-transcendent theory of truth.

The other common (but more reasonable) objection to a pragmatist theory of truth is that it cannot recognize as truth-apt propositions whose truth values were once knowable but are now and forevermore unknowable—propositions like "Julius Caesar slept for five hours on the night before he crossed the Rubicon." But a pragmatic theory of truth simply is not a theory of the truth of propositions of this kind; and I see no reason why we must have a single theory of truth for every type of proposition. We can perfectly well, in keeping with the pragmatic spirit, admit only non-transcendent theories of truth for all types of empirical propositions. That is to say, we recognize propositions like the one just mentioned as truth-apt because we can very well conceive of a context of justification for them.

[20] For an outstanding and detailed discussion of the Peircean theory of truth with respect to propositions of the former kinds, see (Misak 2004).

> Make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible, or are perceived to be possible, for creatures like us. (Flanagan 1991, p. 32)

The conclusions we reach regarding questions about which facts count as reasons for which actions, and which reasons outweigh which others in a given scenario, are constrained by, and must be consistent with, the facts about what can be demanded (at least), and expected (perhaps), from a human agent in that scenario, as determined by the physical, biological and social sciences. Most significantly, these conclusions must be consistent with what we learn about the biological limits, and the psychological, social, and cultural enablers and disablers, of human cooperation.[21]

Let us suppose that, at the ideal limit of inquiry, we have arrived at settled conclusions regarding what can be demanded and expected from human beings in a maximally diverse array of choice situations. These conclusions serve as constraints on any adequate normative theory. But why think that there will ultimately be one normative theory preferred to all others? Because as normative inquirers we must *posit* that only one such set of normative propositions is "fated to be ultimately agreed to by all," and this feature is the final marker of their truth. Jay F. Rosenberg has argued, in a thoroughly pragmatist vein, that having a single, determinate, communal, conceptual scheme for the representation of the world is a non-optional end for a community of rational agents (Rosenberg 1980, ch. 7–9). Rosenberg's focus is on the world as we experience it by means of our senses, and so the conceptual schemes he refers to are scientific theories—theories which are successively replaced by better theories in the course of accommodating the accumulation of seemingly anomalous experiences while simultaneously accounting for the representational successes of the theories being replaced:

> This double accountability of a successor scheme to its predecessor(s)…allows us to make non-vacuous sense of the notion of a diachronic sequence of representational systems tending toward a *limit,* and thus allows us to make sense as well of the notion of a conceptual scheme which embodies an *absolutely correct* representation of the one world. (Rosenberg 1980, pp. 180–181)

Arriving at such a conceptual scheme is what would *count as* reaching the end of inquiry *for us*. This is why developing one is a non-optional end for a community of inquirers. It is the only sort of conclusion to the process of inquiry which we could recognize as the terminus of that process. The pragmatist need not even assume that reaching this end is possible even in principle.[22] That is, he need not assume that, given sufficient, even infinite, time and resources, the path of science would and must end in a single conceptual scheme. Nothing about the ultimate fate of scientific

[21] For some excellent recent discussions of this topic, see (Tomasello 2009; Bowles and Gintis 2011).

[22] Rosenberg claims to show that it is indeed possible in principle; but as far as I can tell, his argument only establishes that it is a non-optional end for a community of agents such as ourselves. Whether it is possible, even in principle or not, is not something that can be determined at the outset.

inquiry need be taken for granted. Rather, the point is that we, as a community of rational inquirers, can have no other goal for our inquiries, other than the development of such a conceptual scheme. This is and must be our aspiration. If it should actually be that no such conclusion is possible, that fact would impart a tragic character to rational inquiry; Peirce's "cheerful hope" would necessarily go unfulfilled. There would then be no single, complete, self-consistent true conception or theory or representation of the world; reality, which is to say reality-as-we-experience-it, would be fundamentally fragmented in something like the way countenanced by Theodor Adorno (1966). But it is at least a tragedy we are spared from suffering, insofar as there is no point at which we would or could conclude that our end is in principle unreachable. For so long as we have not reached it, we can only see ourselves as not yet having completed our task.

Why, though, should we think that *normative* inquiry, and in particular *normative ethical* inquiry, is fit to be modeled in a way analogous to the pragmatist model of empirical inquiry? The suggestion I want to make is that one major aspect of a distinctively human experience of the world is the fact that we experience it as *values-and reasons-laden*, and as populated by other agents who experience it in the same way. To cease to experience the world as a place in which there are reasons to act, or in which there are entities, experiences and states which have value, and in which other agents also experience it in these ways, is to cease to experience it as a human being. If the philosopher's notion of a "faculty of ethical intuition" is to have any place in a modern naturalistic worldview, it is here. So-called intuition is nothing more than the complex of psychological capacities in virtue of which we humans experience the world in these ways. Of course, if we have these capacities, it must be the case that we evolved to have them. The evolution of the evaluative and normative nature of the human experience of the world is a topic I touch on in the next chapter.

Just as a single, scientific representational scheme of the world is a non-optional end for a rational community, so too a single conception of the Normative Order—of what facts are reasons for what actions under what circumstances, and of the weights and scopes of those reasons—is a non-optional end for a community of reason-responsive agents, who cannot avoid experiencing the world as populated by reasons to act, and who cannot escape acting in it on the basis of what they take those reasons to be.[23] Just as the representational schemes of science, moreover, are answerable to sensory experience, our normative scheme is answerable to the indispensable human ends—the general constituents of a good life which we, as humans, *find ourselves with*, just as we find ourselves with our sensory experiences (Rosenberg 1980, p. 175). What I mean by this is not that we acquire knowledge of our indispensable ends non-inferentially, as we do knowledge of what we perceive.[24] Only that we do not exercise control over what these ends are, in the same way that

[23] Though this does not, of course, imply that we necessarily act on what we take to be the overall balance of reasons.

[24] And this, as we know from Sellars' attack on the Myth of the Given, is not to say that the status of perceptual knowledge as knowledge is independent of its inferential connections to other knowledge.

we lack control over our sensory experiences. Our knowledge of which ends are indispensable to a good human life must be an inference made from a broad shared experience of many forms of life, and reflection on that experience aimed at identifying the elements common to those forms of life which we are capable of recognizing, in light of that broad experience, as good.[25]

These ends constrain what could possibly count as an adequate normative theory as well as an adequate conception of the good life, and following Sellars, we can identify the end of living a life that benefits others as the foremost among them: "This commitment to the well-being of our fellow man…stands to the justification of moral principles as the purpose of acquiring the ability to explain and predict stands to the justification of scientific theories" (Sellars 1967, p. 411). By means of our individual and communal/social experiments in living, and our interactions with other individuals and other groups performing different experiments, we come to find one normative scheme, one conception of the web of reasons for action in which we find ourselves enmeshed, inadequate to the task of enabling us to carve out a coherent and sustainable way of life in which these indispensable ends are achieved. We find that one theory asks too much of us, another not enough; that one guides our actions ineffectively or inefficiently, while another does better. We then seek to replace one theory with another which improves on its predecessor while preserving its predecessor's successes.

The idea that the distinctively human experience of the world is an experience of the world as a place laden with values and reasons is an ancient one. It's precursors include Aristotle's notion of *phantasiai*, or value-laden representations of the world; and the Stoic notion of a *phantasia hormetike*, a perceptual experience of the world which itself represents a state of the world as good or bad. As Reeve explains the Aristotelian notion: "Our perceptions and beliefs do not present us with a neutral or value-free world, some parts of which acquire value in our eyes because we already desire them; rather the things they present to us already include elements that perforce instill desire because they are already either pleasant or painful, good or bad" (Reeve 2013, p. 208). In more modern terms, my suggestion is that experiencing the world as shot-through with value and normativity is a necessary aspect of any distinctively human experience of the world, in something like the way Kant believed that experiencing the world as structured by causation was a necessary aspect of any possible experience of the world. Indeed, I believe that meta-ethics as a field has a great deal more to learn from the 1st Critique than from the 2nd, though the latter unfortunately dominates contemporary Kantian ethical (and meta-ethical) theorizing.

We must be exceptionally careful to distinguish the pragmatist view of normativity and value from any form of non-cognitivism, and in particular from Simon Blackburn's projectivist quasi-realism (Blackburn 1993, 2001). The best way to understand the relation between Blackburn's theory and the one being suggested here is by analogy with Hume and Kant on causality. For Hume, the cause-effect relationship is something which is added by the human mind to its experience of the

[25] For an example of some excellent work in this vein, see (Kenny and Kenny 2006).

world, which experience itself contains nothing other than the always-repeated conjunction of events. It is only through the habit of making this addition that we come to take ourselves to be experiencing a world which itself exhibits this relation. By Hume's lights, we can at any moment attend to our actual experience of the world, and recognize that that the causal relation is not present in that experience, but is an addition of the mind. Likewise, for Blackburn, the property of being good or bad, or the reasons-relation between facts, are things which the human mind projects onto the world, and through habit, we take ourselves to be experiencing a world which itself exhibits these properties and relations. On the contrary, my view has it that it is a necessary constituent of any human experience of the world that the world be experienced as a place laden with values and reasons. We cannot—*even conceptually*—separate and isolate some fundamental version of our experience of the world, as a Humean like Blackburn must have us do, and identify it as a value- and reason-free substrate onto which our minds graft normative properties and relations. Any such purported value- and reason-free experience would fail to be a human experience of the world—just as a purported experience of the world without causal structure would fail to be an experience of the world.

Ethical non-cognitivism, even of Blackburn's exceptionally sophisticated variety, lacks the resources to escape the charge of moral relativism.[26] But insofar as Neo-eudaimonism embraces Peirce's variety of Pragmatism, it eludes this problem. The set of normative propositions which constitute the Normative Order at the limit of normative inquiry is the set which is destined to be the subject of permanent consensus (or would be so destined, if we ever could actually reach that limit). There is no threatening sense in which a proposition's membership in the Normative Order is 'up to us.' In particular, there is no suggestion that the normative convictions we come to adopt are based on 'nothing more than' our own emotional responses. We cannot avoid experiencing the world as reasons- and value-laden; and any particular experience of the world as such must of course be an experience of it as populated with particular reasons and values. The content of this experience must come from somewhere, and it may very well be that the *original* source of this concept for any person is his emotional responses to his sensory experiences of the world. But this is not a condition we are stuck with; it is one we out-grow. Our normative judgments are subject to a continual process of revision in the light of critical reflection on our emotional responses, new experiences which cause dissonance with old judgments that do nothing more than express gut reactions, discussion and debate with other members of our community, and discoveries of incoherence among the judgments we make. Normative maturation is the process of retraining ourselves emotionally, so that our emotional responses follow our rationally refined judgments and enable us to be moved to act on them.[27]

[26] For Blackburn's near-heroic attempt, see the last chapter of (Blackburn 2001).

[27] The same can be said of the familial/communal/social rules which are instilled in (most of) us during childhood, and which serve as a bridge between the entirely emotion-based judgments of early childhood and mature ethical judgment.

Normative propositions, moreover, are every bit as capable of being true or false as are the propositions of science, given a pragmatist theory of truth for both. My normative Pragmatism thus joins Realism and Constructivism in the ranks of theories which take normative propositions to be truth-apt, and regard some normative propositions as true. That it is not a form of Realism should be clear: I do not regard true normative propositions as true in virtue of a relation of correspondence with any states of the world. The relationship between my view and Constructivism is more complicated. As I have already promised, in Sect. 4.3 below I will explain the ways in which my Pragmatism differs from any form of Constructivism—differences significant enough to establish it as a rival view alongside the other major meta-ethical theories.

We must be vigilant, however, in maintaining that the conception of the Normative Order, just like the correct scientific representation of the world, is a *limit* concept. Given that the truths of normative ethics are constrained by the truths of scientific inquiry, the final conception of the Normative Order could only be reached at the hypothetical end of inquiry, at the limit of the infinite continuous process of investigation into the natural and social world. Only at this limit could the process of normative inquiry—the process of working out the details of the web of reasons for action in a way consistent with the relevant empirical constraints—itself be concluded. What this means is that the final states of *all three* of the conceptions we have been discussing—*phronesis*, the final state of our conception of practical reason; *eudaimonia*, the final state of our conception of a good life; and the Normative Order, the final state of our conception of the web of reasons for action—*are limit concepts*. *Phronesis* is not complete until it both finishes the process of refining and revising its identification and weighting of reasons for action—the process of constructing the Normative Order—and, drawing on the elements of the Normative Order, completes the construction of *eudaimonia*. *Eudaimonia* is not complete until the practical reasoning constructing it is phronetic. Since the adoption of interests partially determines what reasons for action there are, the Normative Order cannot be complete without our conception of *eudaimonia*—the set of all possible individual complete conceptions of the good life—itself being complete. If we conceive of practical reason, in its ends-deliberation aspect, as a function from the web of reasons for action (along with all the possible outcomes of the process of self-discovery) to fuller and more complete versions of possible conceptions of a good life, the construction of which feeds back into the expanding web of reasons, then we can characterize the limit-point at which both the Normative Order and the complete conception of *eudaimonia* emerge as the point at which this process reaches a final and permanent equilibrium. This is the point at which alterations to the web of reasons for action cease to further alter conceptions of a good life, no further alterations to the web of reasons arise, and so both the set of conceptions of a good life and the web of reasons can finally be said to be complete. This would be the ultimate conclusion of the game of ends-deliberation described in Chap. 4—a final Nash equilibrium reached by all possible ends-deliberators, each of which would have therefore arrived at a final individual deliberational equilibrium. That point is the point at which practical reason itself can finally be said to be complete,

to be genuine *phronesis*. And so *phronesis*, *eudaimonia*, and the Normative Order emerge alongside one another only at the ideal limit of empirical and normative inquiry.

4.2.3 Apparent Reasons, Intersubjective Oughts, and Convergence on Objectivity

We of course cannot wait until we reach this final, ideal consensus before we proceed to act. Indeed, without action, without engaging in a multitude of experiments in living, we can make no progress toward better normative theories. We must lead our lives in the light of our best hypotheses regarding what our reasons are—both our reasons for action, and our evidence for our beliefs and judgments, normative and descriptive alike. What we actually deliberate, choose, act, and evaluate ourselves and others on the basis of are *apparent reasons*, where an apparent reason to believe *p* or do ϕ is a proposition which, if true, would be a reason to believe *p* or do ϕ. None of us can hope to do any better.[28] But what apparent reasons there are for an agent is not simply a function of whatever normative beliefs that agent happens to have. Apparent reasons are *not subjective*, they are *intersubjective*. What apparent reasons there are, for both normative belief and action, depends solely and entirely on what the best available evidence is for both normative belief and action. And the content and confirmation of such propositions of evidence is intersubjective; the determination of what would count as evidence for what, and of what evidence there is, is always a communal endeavor in the normative arena. The prior probabilities of particular agents regarding particular normative claims are subjective. But we have common (and continuously developing) standards for setting both the probabilities assigned to the truth of the propositions of evidence themselves, and to the conditional probabilities that relate that evidence to the claims supported by the evidence. And regardless of what the individual priors are, we expect (insofar as we are Bayesians) that over time, with accumulation of evidence, the updated probabilities assigned by different agents to the normative claims in question, will *converge* (Hawthorne 2004). That which they would ultimately converge on is (for the pragmatist, by definition) the truth. The determination of apparent reasons is

[28] It is pointless to argue, as Kolodny has done *via* a purported analogy with the classic argument for act-utilitarianism over rule-utilitarianism, that we have no reason to follow the evidence in those cases in which the conclusion most supported by all the currently available evidence will in fact turn out to be false. Unlike the act-utilitarian, who can presumably perform the expected utility calculation at any time and determine whether breaking the rule in a given instance is better than following it, we cannot look behind the curtain, as it were, and determine whether our best evidence is, at present, misleading us. If our concept of "ought" is to be at all practical, then, there is no interesting sense in which we ought not follow the balance of all the best presently available evidence, or even in which we ought not have done (in retrospect) when we turn out in the long run to have been wrong. And I simply have no response to Kolodny's completely unfounded doubt that following the canons of rationality has proven itself to be, throughout human history, the most reliable way of approaching the truth. See (Kolodny 2005).

thus a communal endeavor, and is conducted in the light of the best evidence available at the time.[29] And the conclusion of this communal endeavor is the knowledge of what reasons there actually are, which is obtained at the end of inquiry.

Joseph Raz has argued against the claim that any such convergence of normative belief is possible. But his arguments betray a flawed understanding of the notion of convergence. Raz relies here on the importance of *thick concepts*—concepts whose mastery requires knowledge of a shared background of interests and social and cultural practice—in moral thought. For there are many interests, and so many values, which are not universal aspects of the human condition, but are only intelligible in the context of a particular form of life. Raz's first move is to claim that if convergence is supposed to be the convergence of all, then we have good reason to suspect that it is not possible, since it could only be possible for those who are capable of mastering the thick concepts employed by those whose social and cultural backgrounds differ from their own, and not every person is capable of this (Raz 1999b, p. 138). But there is no reason whatsoever to suppose that the sort of Bayesian convergence Peirce writes of, with respect to any domain of knowledge, requires that every person ultimately arrive at the same position, even those who are not capable of mastering the necessary concepts. And it is not at all clear to me why Raz thinks that there is. My best guess is that Raz is conflating the Peircean idea of convergence at the limit of inquiry with the idea, familiar from Rawlsian deliberative Constructivism, that it is possible, at any time, to articulate the conditions of an ideal deliberative situation in which a convergence of normative belief is assured. But as I will discuss shortly, the pragmatism I am advocating is *not* a version of deliberative Constructivism (or, for that matter, of any type of Constructivism).

One might think that the very admission of the importance of thick concepts in moral thought is itself grounds to doubt the possibility of convergence on a single set of normative propositions.[30] For it carries the implication that the truth, or even the intelligibility, of some moral propositions will depend greatly on local sociocultural context. But it is important that we not be misled into thinking that this notion of ideal consensus implies acceptance of a single set of *context-independent* normative truths. It does not. What we accept as the true normative propositions will always be highly context-sensitive and may vary widely between sets of circumstances. This conviction is an expression of my commitment to moral particularism, which I will discuss shortly. The hypothetical ideal moral consensus is a consensus on what normative propositions are true within each possible set of circumstances—or, if we like, a consensus on the truth of a super-set of context-indexed normative propositions.

---

[29] And thus we need not endorse the absurd claim that the man who drinks a glass of petrol, thinking it is a gin and tonic and wanting to drink the latter, has a *reason* to drink the glass of petrol in virtue of his belief and desire. Presumably, he has no business believing that it is a gin and tonic, for any number of reasons—the way it smells, for instance. We can perfectly well explain his action by saying that he mistakenly took himself to have a reason for what he did. He was of course not aware of his mistake; but he may still be held responsible for making it, given the availability of evidence contradicting his belief.

[30] Hilary Putnam, for example, makes this claim (Putnam 1997, p. 169).

Raz does not think it impossible for an individual outside a given socio-cultural tradition to acquire and master the thick concepts employed by those who live within it; he is amenable to the possibility of a Gadamerian "fusion of horizons" (Gadamer 1997, p. 302). So Raz goes on to consider the possibility of convergence strictly among those who are capable of mastering the relevant concepts. His argument against this possibility is based on the fact of "indeterminacy of reasons which may lead rational inquirers, even when they share the same premises, to diverge in their conclusions on those occasions where it would be rational to believe a certain proposition and also rational to doubt it" (Raz 1999b, p. 138). And it is here that Raz betrays the fact that he does not quite understand the notion of Bayesian convergence. It is perfectly true that there may be points in time, in the course of inquiry, at which it is perfectly rational to believe a given proposition, or to doubt it. The whole point of the idea of convergence is that with the passage of time comes the ever increasing accumulation of evidence, and with that accumulation of evidence the probability of one of two mutually exclusive and exhaustive hypotheses will approach 0 while the other approaches 1. In the long run, then, a point is reached in the course of inquiry at which it *ceases* to be the case that the members of a community of rational and expert inquirers can, while remaining perfectly rational, diverge in their conclusions. What Raz intends to be an argument against the possibility of convergence thus turns out to be a flat and unreasoned denial of it.

Raz's purported argument against convergence of normative belief is actually one instance of a general argument strategy—the argument from the ineradicability of 'reasonable moral disagreement.' Of course, there is such a thing as reasonable moral disagreement. It occurs in precisely the sort of situation Raz identifies. But as I have argued, such disagreement is, for the pragmatist, by its very nature temporary; it is not ineradicable (though it may be very long-lasting). Often, what is taken to be an instance of reasonable moral disagreement is not a case of disagreement at all. It is a case in which it appears that two mutually exclusive claims are being put forward, but the appearance of mutual exclusivity is in fact due to the fact that the two claims concern what is right (or wrong) in relevantly different circumstances, the full details of which have not been spelled out. And so what appear to be cases which push us toward the acceptance of ineradicable reasonable moral disagreement, and from there to moral relativism, actually push us toward moral particularism. David Wong, for example, offers many powerful and, to my mind, convincing, arguments that the existence of moral disagreement should push us to accept a view which he calls "moral relativism." Only that view is not in fact moral relativism at all, but rather a form of moral particularism, one in which special attention is paid to the important role of cultural facts as reasons which contribute to the rightness of actions (Wong 2009). What distinguishes the particularist who is sensitive to the normative role of cultural facts from the cultural relativist is, first, that the particularist recognizes the possibility that the members of a given culture hold indefensible views about the normative relevance, weight, or scope of facts about their own practices and form of life; and, second, that there are some types of choice-situations which are normatively invariant under changes in cultural facts. Facts about the specific practices and form of life of a community are not always relevant to ques-

tions of rightness, and there are detectable patterns among the situations in which they are not.[31] The possibility of recognizing the view of the normative significance of the facts of one's culture as indefensible and revising that view, and of recognizing certain types of choice-situations as culture-invariant, are features of all healthy cultures. This distinction between a healthy culture and a sick one is precisely what is unavailable to the cultural relativist.

I do accept that there are cases of ineradicable moral disagreement which are not merely apparent. But I deny that such cases are of reasonable disagreement. Moral convictions are an important component of many socio-culturally shaped personal identities. If having a particular conviction is seen as essential for being counted a member of a particular group, and being counted a member of that group is a core part of an individual's personal identity, it may be the case that no argument, and no amount of evidence contradicting the implications of that conviction, could shake that conviction. An individual may prefer to hold on to the prior conviction—and thus hold on to the group membership and their own established sense of identity—rather than respond to even an overwhelming amount of evidence and force of argument against their position. But such cases are of course *not* cases of *reasonable* disagreement.

One might think that convergence is impossible on the grounds that it is likely impossible even for any one person to master all of the relevant normative concepts. After all, there are simply too many socio-cultural backgrounds of interests and practices for this to be feasible. But in fact, there is no good reason to think, as Raz seems to, that convergence requires that there be even one person capable of mastering all the relevant normative concepts. It is equally implausible to suppose that the final states of the natural and social sciences will be such that any one person could ever be capable of mastering all of the concepts employed in each one. But that itself is no argument against the possibility of convergence in science. In any domain of inquiry, empirical or normative, there must be a division of intellectual labor—the world is far too complex for this not to be needed. The final convergence of belief, in the empirical and normative realms, is a convergence of the beliefs *of a community*. And that convergence consists of an assemblage of smaller convergences within smaller communities—within all those groups who have mastered some single corner of the grand investigative endeavor. It is far too much to expect or insist that the conclusions reached at the limit of inquiry be ones that any one person could possibly have exhaustive first-hand knowledge of.

A different sort of argument against advancing any claims about the convergence of belief is offered by Joseph Heath:

> To speculate about whether judgments will converge is to presuppose that we have some independent access to where our arguments are leading. But this is highly implausible. Even in formal domains like mathematics, often the only way to show that something is provable is to actually prove it. There is seldom any external guarantee that a particular strategy of proof will be successful. Similarly, we do not have any advance knowledge that

[31] "Universal moral principles" are simply the most effective heuristics for dealing with culture-invariant choice-situations.

> physical science will lead to one single account of how the world is. In fact, many philosophers have argued that it will not. But even if this argument were decidable, it is implausible to think that the direction it goes will have any impact on the cognitive status that we attribute to scientific beliefs. (Heath 2001, p. 222)

The crucial point, which I have already made above, is that the pragmatist need not assume that our judgments *will* converge, or even that, given world enough and time, they *would in principle* converge. Rather, he must claim that the members of a community of rational inquirers have no option but to posit such a convergence as the goal toward which their efforts are directed, since there is nothing else that could count for them as the terminus of their inquiries.

But why should we think that implicit in the activity of inquiry is a quest for convergence on a single theory, or perhaps a single coherent family of theories? Progress in any field of inquiry takes two forms. The first is discovery made against a background theory, a set of assumptions which are held fixed at least temporarily. The second is the replacement of one background theory, whose claims have now been placed in jeopardy, with another. As we have already seen, for transitions of this second type to count as progress, the successor theory must be doubly accountable to its predecessor. The successor must account for both the anomalies that provide the impetus to replace the predecessor, and preserve all of the predecessor's successes. If there is progress of this second sort, then, each successor theory is able to do what its predecessor did, and more. As we progress from one theory to another, the body of questions we cannot answer, and problems we cannot solve, shrinks. The limit of this process of progress—the end point to which the process is directed—is the development of a theory that will not need to be replaced, because there are no questions left unanswered, nor problems left unsolved.

Insofar as we commit ourselves to progress in inquiry, then, we commit ourselves at least to striving after convergence:

> To say that an ideal limit to the diachronic process of replacement of predecessor conceptual schemes by qualified successors *exists*…can only be to say that the process itself is, in the specified sense, a *convergent* process—to say, that is, that "the appearances" *are* necessarily continuously saved. And to say that there is but one world will then be simply to affirm that a convergent, nonarbitrary, determinate process of conceptual evolution of this sort is both *possible* and, indeed, for such beings as we in fact are, *non-optional* as well…[T]he claim that there is but one world emerges, surprisingly, as a claim about us—about the kind of beings which we are, and about the kind of conducts which are therefore mandatory for us as a condition of our very existence as beings of that kind. (Rosenberg 1980, p. 187)

What we must be careful not to commit ourselves to is any claim about what must be, or is guaranteed to be, included in the final theory on which inquiry converges—assuming that this limit does exist:

> The concept of a representational system which stands as the limit of such a diachronic process of retrospectively justifiable conceptual scheme replacements is, then, one to which we can give sense. But it is important to add that to thus give sense to the notion of an ideal representational system is not to describe any realizable conceptual scheme, for we can give sense to that notion from our irrevocable perspective as embedded within these evolution-

> ary processes only by invoking the notion of a limit. It is, therefore, a purely regulative ideal… (Rosenberg 1980, p. 186)

This caution, of course, applies to my own views as much as to anyone else's. I cannot claim that the theory of Equal Liberty, or the theory of moral duties and rights to be developed in the next chapter, are fated to be agreed to by all.[32] They are my proposals; they are the contributions I have to offer to the great tradition of moral and political discourse. They rest on numerous assumptions, all of which are open to question. Nor can I make any claims of destiny for my account of the good life, or of indispensable human ends, or of the types of evidence the excellent ends-deliberator relies on. These are background assumptions which my theory holds fixed. They are as answerable to experience and vulnerable to revision as any other claims. But the only way to make progress in moral and political theory is by *doing* moral and political theory. And the only way to make social progress is by *using* moral and political theory. So I, like any other theorist, offer my views backed with the strongest arguments I have to make, rooted in the soundest assumptions I know of.[33]

In characterizing our reasons as apparent, we are simply acknowledging the fact that the normative project is ongoing. We judge how responsive to reasons an agent is based both on what he takes to be his reasons, and on our communal evaluation of his judgments regarding what his reasons are—whether he seems to have based those judgments on the best available evidence for what his reasons are, for example. And we judge the rightness of an agent's action based on the best-supported available view of where the balance of reasons lies in his particular case. We likewise do our level best to determine which ends are worth pursuing, and thus what our interests are, always aware that our conclusions remain ever-revisable. As we increase our evidence, update and refine our views, and strengthen our confidence in our normative conclusions, we can understand ourselves as progressing toward the limit at which permanent consensus is reached, and this idea serves as the regulative ideal that guides our ethical theorizing.

There are profound implications of my claim that, at any given point in time in the midst of the project of normative inquiry, none of us can hope to do better than

---

[32] Even if the just society is the society of Equal Liberty, the policy program I have outlined for its achievement in Chap. 11 may not prove wholly adequate to the task. Perhaps the institutional structure of a society which could fully achieve Equal Liberty would look radically different from what I have described; that possibility cannot be ruled out. But I think there is reason for confidence that that program would make possible the development of agents who were free and autonomous to a high enough degree, so that a transition to a society fully capable of achieving Equal Liberty would be possible. The members of that society would at least be, to borrow a sentiment from Herbert Marcuse, "free to give their own answer" to the question of how society should be structured, and no longer "kept incapable of being autonomous…indoctrinated and manipulated (down to their very instincts)" such that "their answer to this question cannot be taken as their own" (Marcuse 1964, p. 15).

[33] Habermas fails to heed this warning when he attempts to argue that any norm which could be accepted in an ideal speech situation must be one which is equally in the interests of all (Heath 2001, ch. 6).

to deliberate, choose, act, and evaluate ourselves and others on the basis of apparent reasons. Primarily, this means that at any such point, we must define right action *not* as action supported by the balance of reasons *simpliciter*, but as action supported by the balance of apparent reasons—being very careful to remember that apparent reasons are intersubjective. The moral 'ought,' then, is not subjective, and is not primarily objective, but rather is intersubjective. We may still speak intelligibly of an objective ought: we can define those actions which we objectively ought to do as those actions which are supported by the balance of reasons which would be recognized at the limit of normative inquiry. Making such assertions is like asserting that what we ought, objectively, to believe is whatever we would end up concluding is true at the limit of inquiry. Such conceptions of normativity have a role to play as regulative ideals. The point of our efforts is to progress ever closer to the time at which what we believe is what we ought, objectively, to believe, and what we do is what we ought, objectively, to do. But the objective conception is of no immediate practical interest or importance. It has no bearing on the question "What are we to think and do *now*?"

This view of the intersubjective dimension of the normative helps me to hold fast to the claim that requirements of rationality are normative requirements. Cases in which rationality cannot require one to alter one's judgment, decision, or intention—because the sort of evidence in light of which one would do so is inaccessible—are likewise cases in which it cannot be said that one ought, normatively, to do so, at least in the intersubjective sense. For the lack of accessible evidence implies a lack of apparent reasons to do so.

A further implication of this view of the normative is that what we ought to do in a given situation changes over time with the mere progression of normative inquiry. Suppose two agents find themselves faced with a choice between the same two actions at two different times. The circumstances in which they find themselves are identical, except for the epistemic states of the agents: the agent living at the later time has access to more evidence bearing on the normative significance of the features of the situation—evidence which, for historical, social, or cultural reasons, is simply beyond the reach of the earlier agent—but the earlier agent would evaluate that evidence in the same way as the later one if he had access to it. What the agents ought to do, in the intersubjective sense of 'ought,' may be different, though what they ought to do in the objective sense will of course be the same.[34]

Aristotle's characterization of the ideal agent—what he calls the *spudaios*, a sort of *phronimos par excellance*—is perfectly suited to serve as the regulative ideal that guides our understanding of rational and moral agency.

> [T]he *spoudaios* judges each sort of thing correctly, and in each case what is true appears to him…[and *eudaimonia*] is determined by reason and in the way which the *spudaios* would determine it. (*NE* III.4 1113a29, II.6 1106b36, translated in Rosler 2005, pp. 131–132)

[34] Raz argues to a similar conclusion, though from different premises (Raz 1999a, pp. 161–181).

The Aristotelian *spudaios* is actually the image of the agent deliberating and acting at the limit of empirical and normative inquiry. The reasons he sees himself as having, and which he acts on, are not merely apparent. He does what he ought to do in the objective sense, as his inquiries, in co-operation with the inquiries of his fellow *spudaioi*, have taken him to the point where the intersubjective dimension of the normative merges with the objective.

4.3 Neo-eudaimonism and Constructivism

The Aristotelian Pragmatism I have articulated must not be confused with Mark LeBar's very interesting, if ultimately flawed, Aristotelian Constructivism (LeBar 2008).[35] For despite the fact that on my view, our conceptions of *phronesis*, *eudaimonia*, and the Normative Order must be constructed by us in the way I have described, Neo-eudaimonism is not a variety of Constructivism at all. We must construct these conceptions because there is nowhere else for them to come from; they are not "out there in the world" waiting for us to stumble upon them. Neo-eudaimonism, as we have already noted, is not a form of Realism, as this term is usually understood. But our final conceptions are not true *in virtue of* having been constructed, or having been constructed in some particular way. I could not claim any such thing, as I would have to rest such a claim on the special legitimacy of construction *via phronesis*, and my view explicitly states that *phronesis* itself only emerges alongside *eudaimonia* and the Normative Order at the limit of inquiry. The fact that I have characterized valuable ends as ends that an agent could choose on the basis of a well-executed course of ends-deliberation might be taken to suggest that I am a Constructivist about the value of ends. But we can now see that our judg-

[35] LeBar's Constructivism identifies the reasons supporting an action with the contribution performing the action would make toward some agent's leading a life of *eudaimonia*. Since he takes a correct conception of *eudaimonia* to be correct in virtue of having been constructed *via* the exercise of *phronesis*, his view is Constructivist with respect to both *eudaimonia* and reasons for action. LeBar runs into trouble in trying to make his view "Constructivist all the way down," by arguing that the standards for the successful exercise of *phronesis* are themselves constructed—they are in fact constituents of the conception of *eudaimonia* that *phronesis* itself constructs. An agent successfully exercises *phronesis* just in case he "delivers substantively correct judgments about how to live well," where the standard for correctness is accordance with the conception of *eudaimonia*, the content of which is "itself constructed [by *phronesis*]" (LeBar 2008, p. 205). LeBar argues at some length that his view is neither viciously circular nor an objectionable form of relativism, but none of his counter-arguments touch the real heart of the matter. The view allows for unacceptably self-ratifying constructions along the following lines: (1) The agent uses his practical reason to construct a conception of a good life; (2) that conception characterizes a good life as (among other things) a life in which practical reason is used to construct a conception of a good life exactly like it (whatever that may be), thus deeming the exercise of practical reason a success (since no other, non-constructed standard of success is available); (3) having been so deemed makes that exercise of practical reason a success—makes it an exercise of genuine *phronesis*; (4) having been constructed through *phronesis* makes that conception of a good life a genuine conception of *eudaimonia*.

ment of the value of any end, like our judgment of any conception of a good life of which the end is a part, and our judgment of the quality of the deliberation which leads to the end being chosen, is always preliminary and revisable. We use our best current conception of good ends-deliberation to *identify* those ends of which we are currently most confident of being valuable for an agent in a given position. This is the only criterion of identity we have at any given time, and the only one we need. But what *makes it the case* that an end is valuable for an agent in a given position, if it is, is the fact that it will survive to be part of our conception of *eudaimonia* at the limit of inquiry.

Even less, on my view, are the normative truths true *in virtue of* being those normative propositions which would be objects of consensus in some currently specifiable ideal deliberative situation—an original position (à la Rawls), or an ideal speech situation (à la Scanlon). They are true *in virtue of* being the set of normative propositions on which opinion would converge at the hypothetical end of inquiry. I do not even assume that there is some one privileged deliberative situation which is guaranteed to lead to the construction of this set of propositions. The only courses of inquiry capable of leading us to this goal may be diverse patchworks of experiences and methods. And I flatly deny that our reason suffices for the determination of the features of such a situation at the present (or any other particular) time, such that we need only work out what those features would be, in order to position ourselves to determine what the normative truths are right now. This is the dream of Kantian contractualism, but it is a dream purchased at the cost of respecting the relationship between normative truth and the ever-unfolding discovery of the actual, empirical contours and limits of human behavior.

5 Neo-eudaimonism Part II: Contrasting Criteria of Rightness

5.1 Neo-eudaimonism vs. Virtue Ethics

Neo-eudaimonism, despite being broadly Aristotelian, is not a version of virtue ethics—the type of normative theory most readily associated with Aristotle. The minimal requirement for a normative theory to count as a form of virtue ethics is for it to make reference to virtue, or the virtuous person, in its criterion of rightness for actions, as in:

> (V) An action ϕ in circumstance C is right if, and only if, it is the action that would characteristically be performed by a virtuous person in C.

We, however, have defined right action as action favored by the balance of reasons, a criterion which makes no reference to the notion of virtue. The endorsement of a reasons-based criterion for rightness must not be taken as a mark against my claim that Neo-eudaimonism is an essentially Aristotelian theory. Contemporary Aristotle scholars agree that the concepts of "reason for action," "ought," "duty" and

"obligation" are integral to Aristotelian ethics.[36] As Rosler has observed, Aristotle identifies the end of practical reasoning (*phronesis*) with "what ought to be done" (*ti dei prattein*), and identifies what ought to be done with "what reason prescribes" (*ti logon tattei*), or that which is "according to correct reason" (*kata ton orthon logon*) (*NE* VI.10 1143a8–9, IV.1 1119b17–18, cited in Rosler 2005, pp. 133–134). The one who always judges according to correct reason is Aristotle's ideal agent, the *spudaios*, and "in each case what is true [about what is noble and pleasant] appears to him"—i.e. his reason correctly recognizes the reasons for action that apply to him (*NE* III.4 1113a29, translated in Rosler 2005, p. 131). We shall see in the next chapter, moreover, that the notion of virtue does have an important role to play in my theory. It is by developing an account of virtuous action that we will be able to determine the scope of the exclusionary reasons grounded by an individual's interests.

One advantage of Neo-eudaimonism over virtue ethics (and over Kantianism and utilitarianism for that matter), is that it is not *self-effacing*.[37] Let us say that an ethical theory is self-effacing if whatever it claims makes a particular action right, is not what in fact should move the agent to do it, and/or is not what the ethical theory in question says should move the agent to do it. Stocker's case for the self-effacing nature of Kantianism and utilitarianism is well known. virtue ethics is arguably also self-effacing. If I act generously toward my friend, but am moved by the thought that acting in this way makes me generous, and thus virtuous, then I am being moved by the wrong consideration. I should be moved to act generously by the fact that my friend is in need of my aid. According to Neo-eudaimonism, a fact of that sort is both what makes the action right, and what should move the agent to do it. The theory is free of the schizophrenia Stocker attributes to other ethical theories.

5.2 Neo-eudaimonism vs. Moral Perfectionism

Neo-eudaimonism is also distinct from moral perfectionism, at least in its two most commonly encountered forms. The first, human-nature perfectionism, is the view that the good life is the life of developing and exercising those capacities which are essential to human nature.[38] Such a view obviously presupposes a metaphysical theory of human nature and what elements are essential to it. But my view presupposes no such metaphysical theory. I do understand the good life as one which features, among other things, the development and exercise of the agent's physical and mental capacities. But there is no indication that there is some set of such capacities which are essential to human nature, some particular set which one must develop

[36] See (Rosler 2005, pp. 116–177, citing: Everson 1998, p. 10; Louden 1992, p. 34; Irwin 1986, p. 130; Hardie 1980, p. 334).

[37] For the introduction of the idea of self-effacement, see (Stocker 1976).

[38] See, for example, (Hurka 1993).

and exercise in order to live a good life (beyond those required for basic survival and healthy growth and development, which are not by themselves sufficient for a good life), or that a life is good in virtue of being an expression of essential human nature. And I am not convinced that we need make any such assumptions in order to explain why the development and exercise of our capacities is part of any good life. No explanation for this is needed beyond the undeniable fact that as human beings—as the biological, psychological, and social entities which we in fact are—no life which did not have this feature would or could be accepted by us as being a good life.

Nor is my view a version of objective goods perfectionism (the other prominent form), though this sort of view is perhaps a bit closer. Although such views need not commit themselves to a single, exhaustive list of objective human goods, they are distinguished by the fact that they offer some, perhaps short, list of specific goods, the achievement of at least some of which is essential to leading a good life. But despite my commitment to the claim that there are some features of a good human life which are universal—non-optional ends like emotional fulfillment, the development and exercise of one's physical and mental capacities, the achievement of goals and formation of relationships—I do not assume at the outset that there are *any specific* goals, relationships, or activities (beyond those required for basic survival and healthy growth and development) which must figure in one's life if it is to be a good one. Rather, my approach to identifying the good life is process-oriented: any goal, relationship, or activity which could be adopted by an agent as one of his ends through a well-executed course of ends-deliberation counts as a valuable end, and any coherent set of such ends may constitute a valid conception of a good life for that agent.

6 Neo-eudaimonism Part III: Endorsing Moral Particularism

Moral particularism is defined by Jonathan Dancy, its chief proponent, as the view which asserts that:

> [T]he possibility of moral thought and judgment does not depend on the provision of a suitable supply of moral principles. (Dancy 2004, p. 7)

We should, I think, take this to be a statement not about moral judgment in general—as in 'any old' sort of moral judgment—but about good, or capable, or healthy moral judgment. We may define a set of moral principles—being slightly less stringent than Dancy is—as a set of general statements which (a) determine and (b) explain the moral status of most actions; (c) are learnable; and (d) provide a guide to action in a wide range of new cases.[39] Dancy considers particularism to be one

[39] Dancy would say "all actions" and "all new cases" (Dancy 2004, pp. 116–117).

version of what he calls reasons-holism, which is a view characterized by the following two claims:

1. What is a reason in one situation may alter or lose its polarity in another.
2. The way in which the reasons here present combine with each other is not necessarily determinable in any simply additive way. (Dancy 2003, p. 132)

Each of these claims requires some clarification. The first amount to the claim that a fact R which is a reason to ϕ in situation S_1 may, when in situation S_2 in which it is also present, be a reason *not* to ϕ, or a reason to perform some action ψ rather than ϕ, or not be a reason to do anything at all (it may be normatively irrelevant). The meaning of the second claim is somewhat obscure, and Dancy does not do as much as he could to remedy this. It should, I think, be read as a claim about the weight of reasons which is strictly parallel to the first claim about the polarity, or valence, of reasons. Just as reasons do not carry around some fixed valence with respect to actions from one situation to another—a fact that favors an action in one situation need not favor that action in another—reasons do not carry around a fixed weight relative to one another from one situation to another. Suppose that in S_1, R_1 is a reason to ϕ rather than ψ, R_2 is a reason to ψ rather than ϕ, and R_1 is a stronger reason that R_2. Now suppose, as claim (1) tells us is possible, that in situation S_2, R_2 is a reason to ϕ rather than ψ, and R_1 is a reason to ψ rather than ϕ. Claim (2) tells us that, in addition, in S_2, R_2 may be a stronger reason than R_1. Claim (2) should *not* be read as being inconsistent with the claim that the normative support given to a particular action *in a particular situation* can be modeled as the sum of the numeric representations of the weights of the reasons that favor it *in that particular situation*. This point will become important at the end of the next chapter, when I present a formal model of ethical deliberation which is particularist in nature. We will see that a failure to understand it can lead to serious misinterpretation of the nature of particularist ethical reasoning as a whole.

Moral particularism, as I understand it, is consistent with the truth of all the following claims: some individuals do in fact make use of ethical principles in their ethical deliberations, and arrive at sound conclusions by doing so; *conventional* rules are often needed to solve co-ordination problems, such as the rule that everyone in a given country will drive on the right-hand side of the road; *customary* rules, such as rules of etiquette, play an important role in making smooth social interactions possible, by establishing community-wide expectations for what counts as, say, an expression of gratitude; *institutional* rules, such as the No Resource Waste, Liberty, and Harm Principles, may be needed to regulate the deliberations and actions of *institutional persons* (such as the US Senate) and to define the scope and limits of the powers and privileges of *official roles* (such as that of the President of the United States); *legal* rules (or in less developed societies, proto-legal rules) are indispensable to a justifiable and practicable system of coercion and punishment; and ethical rules/principles have a useful and expedient, though dispensable, role to play in moral reasoning as *heuristics*.

I even believe that a staunch particularist can endorse the claim that social rules—both legal and customary—are indispensable to creating and maintaining a

flourishing society, and to leading a flourishing life within a developed society, beyond the functions just noted. That they are indispensable is made clear by recent work in evolutionary game theory. The immense achievements of human civilization are made possible by the fact that human beings cooperate with one another, and show a willingness to share both gains and burdens, on a scale which is unknown in any other species. When we use game theory to model strategic interaction among individuals, we come across certain types of games in which rational but purely self-regarding individuals refuse to cooperate with one another, despite the fact that the benefits that would accrue to them if they were to cooperate are far greater than what they actually attain without cooperating.[40] Rational but purely self-regarding players in such games will either never arrive at a stable equilibrium, or will only ever arrive at equilibria which are sub-optimal.[41] The human propensity to cooperate is explained by the fact that we are not purely self-regarding creatures; rather, we have evolved to possess a rule-conformative disposition.

To understand precisely what this means, we have to introduce the notion of a correlated equilibrium of a game. A correlated equilibrium of a game G is a Nash equilibrium of a game G^+, in which the game G is augmented by an initial move by a player called "the choreographer." The choreographer instructs the other players in the game regarding which pure strategy they should each play. The choreographer directs each player to play the pure strategy which is that player's best response to the other players, assuming that all the other players also follow the choreographer's directive, *from the choreographer's point of view* (Gintis 2010, p. 132). It may not be the pure strategy with the highest payoff, from the individual player's point of view. However, the players are assumed to have a rule-conformative disposition. We can state precisely the extent of an individual's rule-conformative disposition by specifying a number $\alpha > 0$, such that the individual will prefer following the choreographer's directive to any other available strategy so long as the greatest payout from violating the directive is less than the sum of the payout from following the directive and α. If there is no other strategy with a payout great than that of the directed strategy plus α, we say that the player has a *social* preference for playing the directed strategy, despite the fact that it may not be the strategy with the highest payout from the player's own purely self-regarding perspective. If all players follow the choreographer's directives, they thereby arrive at a correlated equilibrium (which may or may not be a Nash equilibrium for G). Social rules—and in a developed society, legal rules in particular—play the role of choreographer. Following them enables groups of interacting individuals to arrive at efficient correlated equilibria which rational but purely self-regarding individuals would never reach.

[40] The so-called Folk Theorem in game theory gives conditions under which cooperation among self-regarding individuals can emerge when there is an infinite series of repeated games. But the assumptions required by the Folk Theorem are very strong, and empirically implausible. They are almost never satisfied in real-world cases of strategic interaction.

[41] An example of the former is the game of Merchants Wares. A one-shot or finitely repeated Prisoners' Dilemma is an example of the latter.

So the creation of social rules, and the evolution of a disposition to follow them, are crucial to human flourishing and the achievements of human civilization. How do these facts square with moral particularism? The key to appreciating their compatibility lies in understanding the way this disposition is modeled. As Gintis puts it, accepted social rules are "*arguments in the preference function that the individual maximizes*" (Gintis 2010, p. 233). What this means is that the individual, when deciding what strategy to play, assigns a bonus positive value to the strategy directed by the choreographer—the one that counts as following the applicable social rule. *Ceteris paribus*, we have an evolved preference for following rules (Gintis 2010, p. 75). The agent in this model is using social rules to partly determine what valuation to give his available strategies. He differs from the purely self-regarding agent insofar as he takes the choreographer's directive to confer value on a strategy above and beyond the instrumental value of the strategy—the value of that strategy's outcome (or the total expected value of its potential outcomes). The fact that we humans do this—that we do not evaluate our available actions solely on the basis of our valuations of their outcomes—is essential to our ability to coordinate our actions and to the achievements that coordination makes possible. But this sort of rule-based non-instrumental reasoning is distinct from ethical deliberation, at least as moral particularists conceive of it.

Ethical deliberation is deliberation about what one ought to do, and about whether one is under a duty to do what one ought. When we treat a social rule as simply conferring additional value on an action—value in addition to the pre-existing instrumental value of that action –we allow that the sum of those values may still fall short of the much greater instrumental value of some other action: the payout of some other strategy may be greater than that of the rule-directed strategy plus α. In such a case, the agent who simply treats rules as conferring additional value on actions will perform the action with the greatest overall value by performing the action with the greatest instrumental value and breaking the rule. If we equate ethical deliberation with the sort of rule-based practical reasoning modeled in the theory of choreographed games, then we will have to say that in a case like this, the agent who performs the action with the greatest overall value and breaks the rule has done as he ought. But ethical deliberation does not work like this. If ethical deliberation is rule-based, then an agent determines what action he ought to perform by determining what action is required of him by the applicable rule. The fact that he places a much higher instrumental value on performing some other action is not the kind of consideration that can make any difference to the question of what he ought to do.[42] When our conclusions about what we ought to do are at odds with the instrumental value we attribute to our available actions, the former serve as a prompt

[42] Some philosophers, such as Heath and Gaus, do follow the evolutionary game-theoretic way of modeling rule-based practical reasoning while equating rule-based practical reasoning with moral reasoning (Heath 2008, ch. 3; Gaus 2011, ch. 7–10). This equivalence between rule-based practical and reasoning ethical deliberation is the basis of Heath's and Gaus' moral institutionalism—the view that morality just is a system of social rules indispensable to coordinating the actions of agents with different preferences. We will return to moral institutionalism at the end of the next chapter, after developing a rigorous model of particularist ethical deliberation and showing how it

to engage in additional ends-deliberation about how we are valuing the potential outcomes of our actions, and thus to revise our judgments of their instrumental value. The difficult problem of developing a rigorous model of particularist ethical deliberation, and of integrating ethical deliberation with the other forms of practical deliberation into a single framework, will occupy us towards the end of the next chapter (after the account of moral duty has been completed). For now, the lesson is that the fact that there is an important sort of practical reasoning that must appeal to rules does not threaten the particularist's claim that healthy moral reasoning need not. These are different aspects of practical reasoning. And although we can identify social rules with reasons for action and treat them as such, the discussion of the next chapter will show that there is good reason not to do so.

The fact that a rule exists and is widely recognized and followed, however, can certainly make a difference to one's ethical deliberations. Given our evolved disposition to conform to rules, the existence of a rule creates an expectation that others will act in a certain way (a way different from how they would be expected to act in the absence of the rule). And one's expectations about the actions of others can certainly make a difference to one's reasoning about what one ought to do. In this way, the fact of a rule's existence can have an effect on the normative landscape. Nonetheless, the ethical deliberator always faces the question of whether the action proscribed by a rule is the action he ought to perform in his particular situation. Expectations about the actions of others that derive from the fact of a rule's existence are one form of input among many for this deliberative process.

Although following social rules in general may be indispensable to human coordination and achievement, it is certainly not the case that there is any one specific set of rules, following which is indispensable. The content of social rules is an object of deliberation, and of more than one sort. On the one hand, a system of social rules is only as good as the system of social ends which it serves. Rules are fundamentally instrumental, even if we must, at times, avoid viewing or treating them that way. The basis for revising, introducing, and extinguishing social rules is an evaluation of their efficiency in enabling the realization of social ends which are themselves worth pursuing. And in determining whether our social ends are worth pursuing, or whether there are not better ones with which they should be replaced, we must look beyond any of our society's rules, and ask ourselves what reasons there are for pursuing them, and whether they really have the value we are attributing to them. Given that humans have evolved a fairly strong rule-conformative disposition, for a given game there may be many possible choreographers (many possible social rules) which would succeed in directing a correlated equilibrium. The realization of that correlated equilibrium is the social end which the rule serves. Evaluating a social rule thus requires reevaluating the correlated equilibrium which it directs members of a society to realize. For the individual, this means deliberating about the value he attributes to his payout in that equilibrium, considering what evidence he has for valuing it as he has been, what reasons he has for pursuing it,

---

can be integrated with the other types of practical reasoning into a single framework, and examine some of the weaknesses in the arguments offered for this view of morality.

and whether he is giving those reasons and evidence appropriate weight. It also means entering into discussion and debate with other members of society about whether they have asked themselves the same questions and deliberated well about how to answer them. The conclusion of this process may be that the current social rule directs them toward a social end which is not worth pursuing, and that another rule needs to be instituted and internalized in order to facilitate arrival at some other reachable correlated equilibrium which is worth pursuing. We shall return to some of these points in the next chapter. On the other hand, even given a non-uniquely optimal system of social rules, one cannot escape the question of whether, on some particular occasion, one ought to follow a rule or not. And here two very different questions must be distinguished. The first is simply whether one ought to do that which the rule tells one to do. The second is whether one ought to treat the rule as a normative constraint, and do what the rule instructs for the reason that the rule instructs it. A staunch moral particularist cannot allow that healthy moral reasoning ever requires that we do the second. He will insist that at best, rules serve as heuristics to be used when time is of the essence.

I should like to have been a staunch moral particularist. But we have come to the one sort of case in which I am compelled to diverge from orthodox particularism. This is the case of legitimately authoritative legal rules. As I will discuss at some length in Chap. 15, the directives of a legitimate authority change the normative landscape in an extraordinary way. (I will have much to say when the time comes on what exactly is meant by "legitimate.") When a legitimate authority directs an agent subject to that authority to ϕ, the fact that the agent has been so directed by that authority *becomes* the agent's reason for ϕ-ing; it *replaces* whatever pre-existing reasons there may have been for the agent to ϕ. And the agent only *respects* the authority insofar as he ϕ's for the reason that he was do directed by the legitimate authority. (Viewing an authoritative directive as simply conferring additional value on the directed course of action does *not* amount to recognizing the directive as authoritative.) Sometimes authoritative directives come in the form of specific orders that a specific person perform a specific action. But the form in which we most frequently encounter them in modern society is as laws, and laws are rules. So the deliberating moral agent who is subject to a legitimate system of laws cannot show the proper respect for those laws without using them as the basis for his deliberations about what he ought to do in situations to which the laws apply.[43] If healthy moral reasoning includes respect for and responsiveness to the legitimate laws one lives under in the situations in which those laws apply, this entails that healthy moral reasoning is not possible, in these cases, without the use of rules—the legitimate laws themselves.

This, then, is the one exception to my commitment to particularism: the particularist's thesis is false whenever an agent, who is subject to a legitimate authority, is

[43] And this is precisely what the judge charged with interpreting and applying statute—the judge whose rulings constitute the most commonly encountered form of specific authoritative directives—is supposed to do, at least from the perspective of the law itself. Whether adjudication normally, or even ever, works that way in actual fact is a separate question.

in a situation to which a legitimate authoritative directive, in the form of a rule, applies. Even in these situations, however, the rules are not normatively fundamental. As we will see in Chap. 15, authoritative directives depend for their legitimacy on being based on the overall balance of pre-existing reasons.[44] Given the requirement that a set of moral principles need only determine and explain the moral status of *most* actions, one might insist that one is a staunch moral particularist *despite* conceding what I have said about legal rules, if one believes that there are sufficiently few legitimate legal rules. I do not think a society of Equal Liberty is compatible with this sort of legal minimalism. The legitimate reach of the law, however, is of course limited, and so a great deal of our everyday ethical decision-making concerns problems on which the law is silent.[45] (Those limitations on the reach of the law are the subject of Chap. 16.) Since I maintain my commitment to particularism within this sphere, I feel justified in describing myself as a *moderate* moral particularist.

Neo-eudaimonism is fully consistent with moderate moral particularism.[46] The primary source of doubt concerning this claim is likely to be the fact that, according to Neo-eudaimonism, human interests are the ground of reasons for action. For does it not follow from this that, no matter what the situation, the fact that a given interest of a given person would be advanced by a given action is always a reason to perform

[44] Gauss objects to views of this type, which see rules (or, more generally, directives) as depending on pre-existing reasons for their authority/legitimacy/validity, on the grounds that such views make it wrong to follow the rules in any situation in which the rule fails to track those reasons (Gauss 2011, p. 139). But as Raz pointed out over a quarter-century ago, this simply does not follow. The directive is a replacement reason so long as the individuals subject to it will, more likely than not, come closer to doing what they ought by following it than they would by trying to determine the balance of reasons themselves. Only in cases of *clear error* ought those subject to a directive to disregard it (Raz 1986, pp. 60–62). Gauss would of course dispute even this; he is committed to a "strong" interpretation of rules, according to which one ought to follow the rules even in clear cases of error or misapplication. But he bases this on a deeply flawed ideal social contract approach to legitimate authority, already discussed in the Chap. 8.

[45] Can customary rules be replacement reasons, as laws can? Presumably they can, so long as they pass the same test that authoritative directives must pass in order to be legitimate: Raz's Normal Justification Thesis, which we will examine more carefully in Chap. 15. I do not believe, however, that there are so many that would pass this test that moderate particularism is threatened. The idea that we should treat much custom as authoritative strikes me as being at the very heart of Burkean conservatism, and thus it may be impossible to be both a Burkean and a particularist.

[46] And as such, Neo-eudaimonism cannot be interpreted as an aggregate maximizing theory—something like the view that what is right is that which maximizes discretely measurable human achievement. For the neo-eudaimonist, *qua* particularist, endorses value-holism as well; and aggregate maximizing ethical theories assume value-atomism. It is the discrete "atoms" of value, whose valence and magnitude remain constant between situations, that such theories aggregate. Neither the satisfaction of interests in commodities, nor in capabilities, nor in functionings is fit to play this role. Though commodities are necessary for every functioning, there is no simple relationship between one's total share of commodities and one's overall level of achieved functioning; the addition of capabilities, past a certain point, makes no contribution to well-being, and may decrease it; and the State's goal of leximining achieved function subject to effort and voluntary risk requires that overall achieved functioning be interpersonally level-comparable, not unit-comparable, as would be required to speak of maximizing aggregate achieved functioning.

that action? It does not. To say that human interests are the ground of reasons for action is simply to say that, whenever there is a reason for an action, that reason is grounded by some interest of some person. Facts about human interests are the *kind* of facts which can be reasons. It is not to say that there is any interest, and any action, such that that interest always grounds a reason to perform that action. Whether or not an interest grounds a reason for someone to act so as to advance that interest in a particular case depends on the particular circumstances of that case. The neo-eudaimonist does claim that the fact that an action will contribute to someone's interest is *normally* a reason to do it. But this does not amount to an assertion that there are defeasible moral principles. Rather, it is an observation of an ethical regularity. There is no *explanatory* role in the theory for even a defeasible principle of the form: *Ceteris paribus*, one has reason to perform an action that contributes to someone's interest. And such principles are entirely superfluous with respect to reasoning about what one ought to do. Nor does the existence of such regularities require an explanation which invokes the existence of moral principles. What regularities there are explained by facts about human life and human society; we are the kind of creatures who cannot lead good lives without regularly acting in ways that advance each other's interests. I will have more to say in the next chapter about the significance of ethical regularities, and the process of determining what reasons there are in particular cases, and what valences and weights they have in those cases.

References

Adorno, T. 1966. *Negative dialectics*. New York: Seabury Press.

Anscombe, G.E.M. 1958. Modern moral philosophy. *Philosophy* 33: 1–19.

Aristotle. *Nicomachean ethics* [*Ethica Nicomachea*], Oxford Classical Texts, ed. L. Bywater. Oxford: Oxford University Press.

Aristotle. *Politics* [*Politica*], Oxford Classical Texts, ed. W.D. Ross. Oxford: Oxford University Press.

Blackburn, S. 1993. *Essays in quasi-realism*. Oxford: Oxford University Press.

Blackburn, S. 2001. *Ruling passions: An essay in practical reasoning*. Oxford: Oxford University Press.

Bowles, S., and H. Gintis. 2011. *The cooperative species: Human reciprocity and its evolution*. Princeton: Princeton University Press.

Broome, J. 2004. Reasons. In *Reason and value: Themes from the moral philosophy of Joseph Raz*, ed. R.J. Wallace, M. Smith, S. Scheffler, and P. Pettit, 28–55. Oxford: Oxford University Press.

Collins, A. 1987. *The nature of mental things*. South Bend: University of Notre Dame Press.

Dancy, J. 1993. *Moral reasons*, 127–143. Oxford: Blackwell.

Dancy, J. 2003. The particularist's progress. In *Moral particularism*, ed. Brad Hooker, 130–157. Oxford: Oxford University Press.

Dancy, J. 2004. *Ethics without principles*. Oxford: Oxford University Press.

Dancy, J. 2005a. Nonnaturalism. In *The Oxford handbook of ethical theory*, ed. D. Copp. Oxford: Oxford University Press.

Dancy, J. 2005b. Should we pass the buck? In *Recent work on intrinsic value*, ed. T. Rønnow-Rasmussen and M.J. Zimmerman, 33–45. Dordrecht: Springer.

Darwall, S. 2006. *The second-person standpoint: Morality, respect and accountability*. Cambridge, MA: Harvard University Press.
Everson, S. 1998. Introduction: Virtue and morality. In *Ethics*, ed. S. Everson. Cambridge: Cambridge University Press.
Flanagan, O. 1991. *Varieties of moral personality*. Cambridge, MA: Harvard University Press.
Gadamer, H.-G. 1997. *Truth and method*. New York: Continuum Press.
Gaus, G. 2011. *The order of public reason*. Cambridge: Cambridge University Press.
Gintis, H. 2010. *The bounds of reason*. Princeton: Princeton University Press.
Hardie, W.F. 1980. *Aristotle's ethical theory*. Oxford: Oxford University Press.
Hawthorne, J. 2004. *A better Bayesian convergence theorem*. Available at http://www.fitelson.org/few/few_04/hawthorne.pdf.
Heath, J. 2001. *Communicative action and rational choice*. Cambridge, MA: MIT Press.
Heath, J. 2008. *Following the rules: Practical reasoning and deontic constraint*. Oxford: Oxford University Press.
Hookway, C. 2000. *Truth, rationality, and pragmatism: Themes from Peirce*. Oxford: Oxford University Press.
Hurka, T. 1993. *Perfectionism*. Oxford: Oxford University Press.
Irwin, T.H. 1986 Aristotle's conception of morality. In *Proceeding of the Boston Area Colloquium in Ancient Philosophy,* ed. J.J. Clearly, vol. I.
Kearns, S., and D. Star. 2008. Reasons: Explanations or evidence? *Ethics* 199: 31–56.
Kearns, S., and D. Star. 2009. Reasons as evidence. *Oxford Studies in Metaethics* 4: 215–242.
Kearns, S., and D. Star. 2013. Reasons, facts-about-evidence, and indirect evidence. *Analytic Philosophy* 54(2): 237–243.
Kenny, A., and C. Kenny. 2006. *Life, liberty and the pursuit of utility*. Charlottesville: Imprint Academic.
Kolodny, N. 2005. Why be rational? *Mind* 114: 509–563.
LeBar, M. 2008. Aristotelian constructivism. *Social Philosophy and Policy* 25(1): 182–214.
Louden, R.B. 1992. *Morality and moral theory*. Oxford: Oxford University Press.
Marcuse, H. 1964. *One-dimensional man: Studies in the ideology of advanced industrial society*. Boston: Beacon.
McBride, M. 2013. Kearns and Star on reasons as evidence. *Analytic Philosophy* 54(2): 229–236.
Mill, J.S. 1863/2002. *Utilitarianism*, ed. G. Sher, 2nd ed. Indianapolis: Hackett.
Mill, J.S. 1869/1978. *On liberty*, ed. E. Rappaport. Indianapolis: Hackett.
Misak, C.J. 2004. *Truth and the end of enquiry*. Oxford: Clarendon Press.
O'Neill, J. 2002. *Ecology, policy and politics*. London: Routledge.
Peirce, C.S. 1894. How to make our ideas clear. In *The collected papers of Charles Sanders Peirce*, vol. 5. Cambridge, MA: Harvard University Press.
Prichard, H.A. 2002. *Moral writings*, ed. J. MacAdam. Oxford: Clarendon Press.
Putnam, H. 1997. James's theory of truth. In *The Cambridge Companion to William James*, ed. Ruth Anna Putnam, 166–185. New York: Cambridge University Press.
Raz, J. 1975. Permissions and supererogation. *American Philosophical Quarterly* 12(1): 161–168.
Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.
Raz, J. 1994. Rights and individual well-being. In *Ethics in the public domain: Essays in the morality of law and politics*, 44–59. Oxford: Clarendon Press.
Raz, J. 1999a. Moral change and social relativism. In *Engaging reason*, 161–181. Oxford: Clarendon Press.
Raz, J. 1999b. Notes on value and objectivity. In *Engaging reason: On the theory of value and action*, 118–160. Oxford: Clarendon Press.
Reeve, C.D.C. 2013. *Aristotle on practical wisdom: Nicomachean ethics VI, Translated with an introduction, analysis, and commentary*. Cambridge, MA: Harvard University Press.
Rosenberg, J.F. 1980. *One world and our knowledge of it: The problematic of realism in post-Kantian perspective*. Boston: D Reidel.

Rosler, A. 2005. *Political authority and obligation in Aristotle*. Oxford: Oxford University Press.
Sellars, W. 1967. Science and ethics. In *Philosophical perspectives*. Atascadero: Ridgeview Publishing.
Sherman, J. *Reason, virtue and supererogation: The unfinished project of 'Saints and Heroes'*. Unpublished MS.
Soames, S. 2010. *What is meaning?* Princeton: Princeton University Press.
Stocker, M. 1976. The schizophrenia of modern ethical theories. *Journal of Philosophy* 73(14): 453–466.
Sumner, W. 1987. *The moral foundation of rights*. New York: Oxford University Press.
Teichmann, R. 2011. *Nature, reason, and the good life: Ethics for human beings*. New York: Oxford University Press.
Tomasello, M. 2009. *Why we cooperate*. Cambridge, MA: MIT Press.
Urmson, J.O. 1958. Saints and heroes. In *Essays in moral philosophy*, ed. A. Melden. Seattle: U of Washington Press.
Williams, B. 1981. Internal and external reasons. In *Moral luck*, 101–113. Cambridge: Cambridge University Press.
Wilson, J.C. 1926. *Statement and inference with other philosophical papers*, ed. A.S.L. Farquharson, 2 vol. Oxford: Clarendon Press.
Wong, D. 2009. *Natural moralities: A defense of pluralistic relativism*. Oxford: Oxford University Press.

Chapter 14
From Moral Duties to Moral Rights

1 Introduction

We now have a thorough understanding of the nature of natural moral duty—moral duty that is grounded by the interests of others, and in the absence of a directive issued by a legitimate authority. In this chapter, I finally answer Anscombe's challenge by developing the normative ethical component of Neo-eudaimonism, which explains how, and under what circumstances, the interests of individuals ground moral duties, and how moral duties ground moral rights. I then complete the sketch of Neo-eudaimonism by presenting a comprehensive theory of practical reasoning, which reveals the roles that recognition of what one morally ought to do, what one has a moral duty to do, and what others have a moral right that one do, play in ethical deliberation, and which relates ethical deliberation to the other types of deliberation discussed in earlier chapters. But first, we must get clear on what is meant by a moral right, so that the relationship between duties and rights will be intelligible.

2 The Concept of a Moral Right

My discussion of rights is limited to *moral claim-rights*. Person A has a moral claim-right that person B ϕ if, and only if, (1) B has a moral duty (or obligation) to A to ϕ and (2) B's duty to ϕ is justifiably enforceable. For the duty to be justifiably enforceable means that there is (a) some attainable state of the world in which there would be (b) some morally justifiable way for (c) some person or group to enforce the duty. Condition (2) is required because, as Matthew Kramer points out in his exegesis of Hohfeld's jural relations, "A genuine right or claim is enforceable" (Kramer et al. 2000, p. 9). That claim-rights are justifiably enforceable is what distinguishes them from mere demands, even demands that someone do what he in fact

A portion of this chapter has appeared in print as (Sherman 2010).

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_14

has a duty to do.[1] If this is correct (and I am inclined to think that it is), we cannot say that one person has a moral right against another unless, were that right violated, it would be possible for the right to be justifiably enforced.

One need not, however, endorse this conceptual thesis in order to hold the view that we *should not* say that someone has a moral right unless that person's claim is justifiably enforceable. There is a moral argument, recently articulated by Raymond Geuss and further developed by Susan James, for restricting our attribution of rights in this way (Geuss 2001; James 2003). The basic premise of the argument is that "rights are best understood as practical entitlements which make a difference to the lives of those who hold them" (James 2003, p. 133). If that is the best way to understand rights, then to insist that some person or group has a right to receive some benefit when there is no way for that purported right to be justifiably enforced is empty rhetoric; and in many cases it amounts to "a bitter mockery of the poor and needy" (O'Neill 1996, p. 133). Enforceable rights are the only sort that are capable of making a real practical difference to people's lives. And there is something morally and politically perilous about claiming that someone has a right whenever there is some benefit that we think he ought to be provided with (Geuss 2001, p. 146). Doing so allows us to insist that we recognize the basic interests of all people, and this sounds like some sort of moral achievement. But this practice may distract us from two morally significant facts. First, the political systems under which many people live fail to secure for their citizens the possession of enforceable rights, even while they declare that they recognize those rights. Second, in some cases there is no one who could plausibly be identified as the bearer of an enforceable duty to provide another person with some important benefit. In these cases, claiming that those in need are right-holders does nothing to bring us closer to a social arrangement in which their needs are met. Those who are unconvinced by the Hohfeldian conceptual claim or by this moral argument may take me to assume that moral rights are justifiably enforceable, and consider my argument to incorporate that assumption.

The claim-right and the duty that can be justifiably coerced are correlative: one exists if, and only if, the other does. A claim-right is thus always a right held *against* some distinct person or group, the bearer of a duty to the right-holder. A few points are worth noting. First, this definition is neutral with respect to the explanatory relation between correlative rights and duties. It does not imply that we have duties in virtue of the rights of others, or *vice versa*. Second, many traditional 'rights' do not fall within my use of the term, most notably the civil liberties (though rights that others not interfere with one's exercise of one's liberties are included). Finally, not every moral duty need have a correlative right. Only those duties one can be

[1] There is a related debate about the nature of legal rights, between the *will* theory and the *interest* theory, over whether an individual must *herself* have the power to enforce her claims in order to count as a right-holder (as the will theory holds); or must only benefit (in certain ways or under certain circumstances that need to be spelled out) from the enforcement of his claim by *someone* with the required power in order to count as a right-holder (as the interest theory claims). I explore this separate but related debate in my "Dialectical Deadlock and the Function of Legal Rights" (Unpublished MS).

justifiably coerced into fulfilling do. I will occasionally refer to such duties simply as "correlative duties."

3 Neo-eudaimonism Part IV: Interests, Virtues, Duties and Rights

My claim that interests can ground duties can be broken up into the following subclaims: (a) some interests ground protected reasons; (b) these protected reasons are normally conclusive (i.e. their first-order components normally outweigh competing unexcluded reasons); and (c) frequently, all the competing reasons which are present fall within the exclusionary scope of that conclusive protected reason, making it a pre-emptive reason (a duty). And so I will first argue that interests with certain features ground first-order reasons that normally outweigh competing reasons. Even when these interests fail to ground duties, contributing to them is still, normally, the right thing to do. Second, I will argue that those interests also ground exclusionary reasons that capture some common types of competing reasons. When the exclusionary reason grounded by one of these interests manages to exclude all the actually present competing reasons—that is, when the present competing reasons all fall within the exclusionary scope of the reason grounded by the interest—that interest grounds a duty.

The first two of the features in question are *importance* and *relevance*, and I shall explain precisely what I mean by these in due course. The duty that can be grounded by an interest with one or both of these features is imperfect. This means that the scope of the exclusionary reason the interest grounds is not wide enough to capture every possible competing reason. An agent might have an imperfect duty at some time t_1, at which time all present reasons against performing a given action are excluded; and yet, due to changes in circumstances, he may cease to have that duty at time t_2, for at that later time, some other competing reasons may have arisen which lie outside the scope of the exclusionary reason which partially constituted the duty at t_1. Such a duty, so long as it persisted, would be imperfect. And should it be extinguished, the action which was formerly a requirement of duty would become supererogatory. Important interests ground *undirected* imperfect duties, while relevant interests ground *directed* imperfect duties; we shall see what I mean by these terms presently. Of course, an interest may be both important and relevant, in which case it can ground a duty stronger than the one grounded by an interest with only one of these features. The reasons grounded on such interests normally outweigh additional types of competing reasons, and have a wider exclusionary scope. Perfect duties are stronger still. They are grounded by interests which are not only important and relevant, but whose satisfaction is necessary to the flourishing of the society in which the interest-holder lives. The reasons grounded by these interests have the widest possible exclusionary scope. But as we shall see, only a subset of these manage to ground moral rights.

3.1 Undirected Imperfect Duties

The first feature in virtue of which an interest can ground a duty is importance to the interest-holder. My goal in this section is to give some content to the notion of importance, by providing criteria for an interest to ground a pre-emptive reason with an exclusionary scope wide enough to capture some common types of competing reasons, and a first-order component strong enough to normally outweigh those that are left unexcluded. Raz's account offers little help here, telling us only that important reasons are especially valuable to the interest-holder.

A plausible place to begin is with interests whose satisfaction is *necessary* to the interest-holder's pursuit of his ends. Given that our resources are finite, a reason to advance one person's well-being often competes with other good reasons for action. These competing considerations may take the form of agent-neutral reasons to advance the interests of some other person, agent-relative reasons to advance the interests of oneself or of those close to one, or obligations to perform some other action. Suppose that A and B each have some interest, that I am capable of advancing either but not both, and that each interest grounds only an agent-neutral reason for me to advance it (we will hold off on involving agent-relativity for the moment). I want to know which one I have more reason to advance. Now suppose that satisfying A's interest, by ϕ-ing, is *required* for his (continued) pursuit of one of his *central* ends, while satisfying B's interest, by ψ-ing, will contribute to one of his central ends but is not necessary to his pursuit or achievement of it. By central ends, I mean those ends that are high-ranking and whose pursuit is a long-term project. This is a start, but B's interest could still very well ground a stronger reason for me than A's. A's interest could be one which he should satisfy for himself. This will be so if A can ϕ and either (1) A has an interest in *A's* ϕ-ing (over and above his interest in someone ϕ-ing), or (2) ϕ-ing would involve no significant sacrifice for A. So let us further assume that A cannot ϕ or that ϕ-ing would be a significant sacrifice for him. The next question is whether it is A's fault that he is in this position, whether his inability to satisfy this interest is due to his own recklessness or negligence. If not, then A's interest will normally outweigh B's. If it is A's fault, then the last thing we need to know is whether this is an interest which must be satisfied if A is to lead *any* sort of valuable life, regardless of whether that life accords with his own specific plans and goals. This would be an interest in the bare necessities for leading a worthwhile life, and would again normally outweigh B's interest.

So we can draw a preliminary conclusion. An important interest is one which (1) (a) is necessary to the interest-holder's (continued) pursuit of at least one of his central ends, (b) the interest-holder cannot satisfy for himself without significant sacrifice and (c) the interest-holder's inability to satisfy is not his fault; or (2) is an interest in (a) the bare necessities for leading a worthwhile life which (b) the interest-holder cannot satisfy for himself. To these two I must add a third: if satisfying A's interest is necessary to his well-being and A *should* satisfy it for himself, then A has an important interest in not being prevented or discouraged from doing so even if B's interest could be advanced by such interference. In this case too, the reason

grounded on A's interest will normally outweigh the one grounded on B's. I persist in claiming that if A's interest is important and B's is not, A's interest will only *normally* outweigh B's. B's interest might ground an agent-relative reason for me, or the common good might be advanced by the satisfaction of B's interest more than by the satisfaction of A's *despite* the lesser impact on B's own well-being. I address these cases below.

We now have to ask whether only interests that are necessary to the interest-holder's (continued) pursuit of his current central ends meet the definition of an important interest. It seems not. In addition to one's current projects, there are the numerous other valuable projects that one could incorporate into one's life, whether at present or in the future, and whether in addition to or in place of one's current projects. So long as one has not incorporated a project into one's life, one cannot be said to have an interest in achieving its end, and one's interests will not ground reasons for others to help one achieve that end. But provided that it is a project one could incorporate into one's life, one does have an interest in not being prevented or discouraged by others from choosing to pursue it, and in being encouraged to develop the ability to choose for oneself in the first place. This is just one's interest in being free to determine which valuable ends to dedicate oneself to, in making that determination for oneself, and in being free to pursue the new ends chosen on the basis of that determination. It is, in other words, one's interest in liberty, in leading a life of freedom and autonomy. This too is an important interest. Normally, if one of B's interests could be advanced by violating this interest of A's, the reason to advance B's interest will be outweighed. The interest in being at liberty to determine and pursue one's own valuable ends thus joins the ranks of important interests.

Important interests, I will argue, ground exclusionary reasons in addition to first-order reasons. That an interest is necessary to the pursuit or achievement of someone's central ends is a reason not to act on some types of reasons that compete with the reason to advance that interest. More specifically, if I am in a position in which I must choose between advancing one person's important interest, and satisfying an interest of someone else (say myself, or someone close to me) in some *peripheral* end, I would do wrong in ignoring the important interest. By a peripheral end, I mean one that is low-ranking or whose pursuit is only a short-term project. The reasons for satisfying the peripheral interest in this case are excluded, and I am thus under a duty to act to advance the important interest. The duties grounded by the important interests of others, however, are imperfect. For if we change the situation, so that my alternatives are to advance one person's important interest or contribute to one of the more central ends of someone else (again, say myself or someone close to me), then although advancing the important interest may be what I ought to do all things considered, the reasons for omitting this action are no longer excluded. I am morally free to advance my own central interests or the central interests of those close to me, even if these interests lack the feature of importance—even if, that is, my contribution on this occasion is not *required* for the achievement of those ends.

To provide an argument in support of these claims, I must incorporate the notion of virtue into my account. I will therefore pause to describe the essential features of

a neo-eudaimonistic conception of virtue, including the points on which this conception differs from the traditional Aristotelian conception of virtue, as well as my reasons for being unmoved by some recent critiques of the use of the concept of virtue in moral theory in general.

3.1.1 Interlude: A Neo-eudaimonistic Account of Virtue

I begin with a claim about the relationship between right action and the good life. It is possible for an agent always to do what is right, and yet fail to live a good life—at least in the robust sense of "a good life" in which I have been using this term. The agent's circumstances may simply make it impossible for him to live a good life. The fact that, within these strained and desperate circumstances, the agent always manages to perform the action which is, in context, most supported by reasons, does nothing to obviate this unfortunate truth. In my characterization of a good life as a life of freedom, we already have a solid basis for accepting this claim. What reasons for action one has, and thus which of one's available actions counts as right, depends on what one is free to do, as well as on what one is free to become and to be (i.e. what abilities one is free to develop and exercise). One can, in every instance, do what is right within the confines of one's circumstances—including circumstances of severely limited freedom—without living a good life in any plausible sense.

The feature I add at this point to my characterization of a good life is that it is a life thick with instances of a wide range of virtuous actions. This means that a good life, at least in the fullest sense, must be lived in the sort of happy circumstances in which virtuous actions are possible for the agent, and are frequently the actions which the agent ought to perform. To put this point in explicitly Aristotelian terms, I am claiming that a good life must to some extent be a life of *blessedness*, and that this is so because some degree of blessedness is required for there to be sufficient opportunities to perform virtuous actions. Aristotle does not hold this thesis. In his discussion of blessedness, he asserts that a man can be virtuous even though he is cursed, and that his nobility will shine through his adverse circumstances (*NE* I.10 1100b12–1101a8). I do not dispute that this is so in the case of some virtues (fortitude most obviously). But Aristotle believes that *eudaimonia*, and not just the regular exercise of a few of the virtues, can be achieved in any circumstances, no matter how disadvantageous. This is the claim I dispute. I should note, however, that the distance between Aristotle and myself on this point is not so great as it might seem. Aristotle acknowledges that "no function of man has so much permanence as excellent activities…and of these the most valuable are more durable because *those who are blessed spend their life most readily and most continuously in these* [emphasis added]" (*NE* I.10 1100b12–16, translated in Broadie and Rowe 2002). So my claim is a stronger version of one that Aristotle does accept.

But what precisely do I mean when I speak of "virtuous actions"? Here I must draw a distinction between *prima facie* virtuous action and virtuous action *simpliciter*. A *prima facie* virtuous action is any action which is aptly characterized by a

virtue term. When we characterize an action by a virtue term, we are saying that the action was done for a certain kind of reason—the kind of reason that the virtue term tracks. An agent acts (*prima facie*) generously, for example, when he gives some of what he has to another *for the reason that* the other has need of it. The fact of the other's need is the reason for which the agent acts, and so the agent's reason for action is the sort of reason that the relevant virtue term tracks. Someone who was acting *qua* utilitarian, on the other hand, might aid the person in need because he has determined that this action would most increase overall utility, but would instead aid someone already well-off if he thought that *that* would most increase overall utility. So his reason for action is not the sort of reason tracked by the virtue term "generous." An action is virtuous *simpliciter* when it is both *prima facie* virtuous and supported by the balance of reasons (so it is also right).[2] *Prima facie* virtuous actions are a class of actions which are in fact normally right. They are the sort of actions which must be done regularly by most of us in order to establish and maintain a just society and harmonious communities. Thus the traditional connection between the virtues and moral education—to be taught to act virtuously is, *ceteris paribus*, to be taught to be a good citizen and a good neighbor. And the fact that such actions are normally right explains the expediency of virtue-based rules ("do what is generous," or "act as a generous person would act") as moral heuristics—generally reliable short-cuts to conclusions about what one ought to do in a variety of situations.

But which virtue terms do we recognize as legitimate, as picking out "really" virtuous actions? On this point, I can to some extent fall back on my commitment to particularism, and assert that the determination of which actions count as (*prima facie*) virtuous depends to a considerable extent on historical and socio-cultural factors. From a historical point of view, we must be sensitive to how much knowledge of human nature, and of the possibilities and limitations of human society, the denizens of a given historical epoch could reasonably be expected to have. And from a socio-cultural point of view, we must understand that what actions count as virtuous within given community will and must be shaped to some extent by the particular parochial interests (to use Raz's term) of that community. We should not, however, be too hasty to conclude that there are no universal virtues. It is a central tenet within positive psychology that there is a small set of core, universal (or near universal) human virtues. This commitment is founded on the recurrence of certain virtues in nearly all major systems of ethical and religious thought across eras and cultures. These include courage, temperance, generosity justice, friendship and wisdom (Peterson and Seligman 2004, ch. 2). Nor should we be too quick to accept a characterization of a type of action as virtuous simply because it is held to be so in some culture. It is perfectly possible for any given system of cultural belief to be

[2] The only way to ensure that one's actions will be virtuous *simpliciter* is thus to develop *phronesis*, and in particular, its aspect as excellence in ethical deliberation (which we shall examine closely at the end of the chapter). The excellent ethical deliberator is able to see past the surface of a situation, and recognize, for example, that an action which appeared at first glance to be generous is in this instance, owing to the particular facts of the case, not generous but wasteful.

wrong about what counts as virtuous, even in the context of that culture. Cultural history is a complex and highly contingent matter. Cultures may develop in which, for example, extreme suspicion and distrust are seen as praiseworthy, and the one who practices them, laudably careful and shrewd in his dealings with others.[3] But any society whose members adhered to such cultural norms would be an impoverished one. Deep and constant distrust of one's neighbors is incompatible with leading a productive, creative and flourishing life. The culture that identified such actions as virtuous would be a sick one, and as we have seen. the particularist distinguishes himself sharply from the cultural relativist in acknowledging the possibility of sick cultures.

We must therefore reinterpret the question posed above as "Which virtue terms should *we* recognize as picking out genuinely virtuous actions," where "we" are those who are interested in creating and maintaining a modern, just, liberal society and political community (and a society of Equal Liberty in particular). Accomplishing this goal requires not only good institutions and good social and customary norms; it requires that the members of the body politic be possessed of liberal virtues, that they have the capacity to appreciate, respect, and preserve these elements of their society.[4] The virtues we should recognize are those most conducive to meeting this goal. William Galston has argued persuasively that there is a set of distinctively liberal virtues, in addition to the core virtues discussed above which transcend differences of social and political organization. These include independence, tolerance, creativity, initiative, reliability, civility, adaptability, respect, leadership, open-mindedness, and a willingness to recognize flaws in the social system and resist hypocrisy. Devising and arguing for an exhaustive list of virtues—or even of liberal virtues—is no part of my task here, and indeed, I know of no sound, principled way of going about it. I am happy to endorse Galston's treatment of the subject in general, as well as the inclusion of the virtues he identifies on any list of liberal virtues. The importance of one or two additional virtues, not mentioned by Galston, to a liberal society will become clear in what follows.

I have tried thus far to limit myself to a discussion of virtuous actions. But "being virtuous" is not only thought of as a feature of actions. Traditionally, it is also (and in fact primarily) thought of as a feature of agents. Since the ethical theory I adhere to is not a form of virtue ethics, I do not take the virtuous agent as fundamental, and then define virtuous (and thence right) action by reference to him. What I wish to

[3] An extreme example: The Imbonggu tribe of New Guinea distrust the members of neighboring tribes to the extent that they will not accept food from them, for fear that it contains poison (Wormsley 1993, p. 47).

[4] It is a mistake to think that, in addition to good institutions, a just society requires *either* a good system of supporting social and customary norms *or* a virtuous citizenry, as if these latter two had some difficulty coexisting. Any set of official institutions (including a system of laws), will require the support of a less formal, more flexible system of customary norms for its stability and smooth operation. But *both* of these—institutional and customary norms—are of little use if they have no backing but the threat of coercion or censure (respectively) amidst a vicious populace. Virtue is what enables us to determine, in particular cases, that we ought to act as the norm directs us to, and to follow through on that determination even when doing so is difficult.

do, rather, is define the virtuous agent in terms of my reason-tracking account of virtuous action. More precisely, a virtuous agent of a particular type—for example, a generous agent—is an agent that can accurately be described as having, or possessing, the relevant virtue. And we are entitled to say of an agent that he has a virtue, just in case two conditions are met. First, the agent's life is thick with instances of him performing the relevant type of virtuous action. No one is generous who does not regularly act generously. Second, the agent's capacity for self-control with respect to disabling or interfering environmental cues concerning the virtuous action is well-developed. I discussed the psychological phenomenon of self-control at some length in Chap. 7. Recall that there is a particular psychological mechanism, ego-depletion, whereby an agent's capacity for self-control is diminished. We saw that there is ample support in the psychological literature for the claims that individuals can work to make themselves less susceptible to ego-depletion, and that the structure of the choice situations one encounters can make one more or less susceptible to ego-depletion. Becoming virtuous, then, is to a certain extent a matter of the agent's own effort—a matter of moral exercise. But it is not *solely* a matter of personal effort. The very possibility of virtuous action depends to some extent on the way frequently encountered choice situations are structured within one's society. One can only become a virtuous agent within a social context which is conducive to the development and exercise of self-control, a context in which the occurrence of ego-depleting choice situations is minimized. And being virtuous is not simply a matter of maintaining one's self-control in the face of adverse circumstances. It is also a matter of navigating the landscape of potential choice situations one might encounter to minimize opportunities for ego-depletion; it is, as Nancy Snow terms it, a form of "social intelligence" (Snow 2009).

In characterizing virtuous agents as I have done, I have not assumed the existence of robust character traits, these being the bugbear of the so-called "situationist critique" of the use of the concept of virtue in moral theorizing. I do not assume, in saying of someone that he has the virtue of generosity, that there are no circumstances in which he will fail to be generous—will fail to overcome the temptation to act selfishly or callously. To be fair, the apparent force of the situationist critique does not come from the mundane fact that we can find situations in which individuals fail to act virtuously. Rather, it comes from the alleged discoveries of contemporary social psychology that the behavior of psychologically healthy agents can be greatly altered by "subtle situational forces" which are of no moral, and, from the point of view of common sense, very little motivational, significance (Sabini and Silver 2005). More precisely, the situationist critique can be characterized as follows:

> morally significant behavior is affected by features of [the agent's] immediate situation which (1) are not in themselves of moral significance, (2) are not of great motivational significance, (3) are not well known either to laypeople or to the philosophical literature, (4) are numerous, and (5) do not form a coherent class from the point of view of folk psychology. (Sabini and Silver 2005, p. 561)

But as the psychologists John Sabini and Maury Silver have forcefully argued, the philosophers who have drawn this conclusion from the available social psychological research have passed well beyond what the actual evidence supports.[5] Only the first part of this conclusion is warranted. The morally insignificant situational features which affect morally significant behavior are the anxiety that accompanies the awareness that one's view of the present situation—whether of its physical or its ethical features—differs from one's perception of the views of other capable agents when experiencing the same or similar situations, and the embarrassment that accompanies any thoughts of confronting those other agents with one's own view (Sabini and Silver, p. 561). This anxiety and embarrassment are indeed morally irrelevant, so long as what we mean by that is that being out of step with the views of others does not in any way *make it the case* that one's view is wrong. They are, on the other hand, relevant to moral *reasoning*, if only as heuristics; indeed, it is the usefulness in many cases of precisely this type of anxiety as a heuristic in moral reasoning which *accounts* for the powerful influence it has on behavior. Sabini and Silver are willing to grant that these features are not, from the perspective of common sense, of very great motivational significance. I disagree with this point entirely. Anxiety and embarrassment about "not fitting in" and looking foolish in front of one's peers is one of the most significant sources of motivation, from the perspectives of both common sense and behavioral psychology. But they are quite adamant that the evidence does not support the final three parts of the critique. These forces are familiar to everyone, and far from being numerous and incoherent, they are a tightly bound few.

The trouble for situationism, however, extends well beyond problems with the interpretation of psychological research. There are problems with the structure of much of this research itself. The sorts of experiments which the critique relies on are designed to elicit inconsistencies in behavior through the manipulation of morally insignificant situational features. And so the alternative situations in which human behavior is studied are meant to be similar in all morally significant respects. But there is a significant problem lurking in this very experimental framework. We must ask: *Similar according to whom?* As Nancy Snow has argued, drawing on the research of the social psychologists Walter Mischel and Yuichi Shoda, when individuals are observed across a wide range of *subjectively similar* situations—that it, situations which are similar given the point of view, value system, and experience *of the experimental subject*—moral behavior is found to be *remarkably consistent* (Snow 2009). This is what we would expect to observe if, *contra situationism*, individuals did possess robust and stable character traits. Significant situational variation is observed across so-called "objectively similar" situations, which may indeed be morally objectively similar if designed by experimenters of great moral sensitivity, but which are just as likely to be *merely subjectively* similar *from the point of view of the experimenter*.

I would like to register one further objection to the situationist critique from the perspective of my own view. The philosophers who have put forward the situationist

[5] The *loci classici* for the situationist critique are (Harman 1999; Doris 2002).

critique have drawn on one body of literature within social psychology while ignoring another. What they have ignored is the literature on self-control and ego-depletion discussed in Chap. 7. The sorts of situations on which the critique is based—situations which, as Sabini and Silver identify, are conducive to a distinctive sort of social anxiety and embarrassment—are just the sort of situations in which a great deal of self-control is required in order to follow through on one's moral judgments. In particular, more self-control is required than anyone should expect to have who has not made a special effort to develop and enhance his powers of self-control. The question which requires further research, then, is whether there are any feasibly employable methods of developing one's self-control which will specifically enhance one's ability to cope with social pressure to alter one's judgments, or to act contrary to them. If it turns out (as I suspect it will) that there are, then the death knell for the situationist critique will have been tolled.

Although I have given an account of the virtuous agent which does not appeal to robust character traits—relying instead on reasons-tracking account of virtuous action and on the empirically well-established psychological capacity for self-control—my doubts about the situationist critique leave me comfortable with the idea that individuals do have robust character traits. In fact, for all the evidence actually shows, it may be that the entire core of Aristotle's account of virtue survives: virtue as a (1) stable disposition (2) to choose an action, (3) with the knowledge of what is being chosen, (4) on the basis of practical reasoning (5) of the sort that a practically wise person would carry out (*NE* II.6 1105a28–1107a2). There are, however, three further aspects of Aristotle's account with which I do not wish to be burdened. The first is the idea that virtue is, in some sense, a mean. I do not wish to deny that many virtuous actions (both the *prima facie* and the *simpliciter*) avoid, in some intelligible sense, excess and defect. Nor do I wish to deny that moral emotion—how one feels about one's practical conclusions and one's actions—is an important dimension of moral life, and that the moral emotions of a good person can similarly often be characterized as avoiding excess and defect.[6] I only wish to deny that we are under some compunction to find an intelligible way to characterize in this way *every* type of virtuous action we might wish to recognize, or that attempting to do so is always a fruitful exercise. I do not feel compelled to accept the claim that gives birth to this compunction, viz., that "being a mean" is an essential feature of anything we rightly recognize as a virtue. I will, however, hedge my doubts about the "Doctrine of the Mean" in the following way. As we are about to see, the primary function of the concept of a virtue in my account is to set the scope of exclusionary reasons in particular circumstances. So I am willing to *reinterpret* the Doctrine of the Mean in terms of appropriate exclusionary scope. The excessive agent acts as if he has exclusionary reasons with wider scopes than they actually have; the deficient agent, narrower scopes. The virtuous agent acts in accordance with the correct scopes of his exclusionary reasons. So by "virtue is a mean," we should understand something like the fact that concepts of the virtues guide us in

[6] I am inclined to agree with Hardie that this is "an important part of the truth about moral goodness" (Hardie 1977, p. 35).

setting the scope of our exclusionary reasons in a way that is neither too narrow nor too wide.

The second point I wish to deny is that there is any important distinction between the virtuous agent and the *enkratic*, or continent, one. But we should be clear about what it is that is being rejected. Aristotle does *not* claim that the virtuous agent does what is right *effortlessly* (Hardie 1977, pp. 44–45). A continent agent struggles to pursue a good end rather than a bad one. A virtuous agent may also struggle in certain circumstances, though not in this way. Rather the good ends of a virtuous agent may conflict, and it may require a great effort to make the choice he determines is right.[7] This problem is particularly acute when one with a great deal to live for is called to have the courage to choose death. The thesis I reject, then, is that in order to count as virtuous, one must never need to struggle to avoid pursuing what one recognizes as bad. This is too high a bar; it violates, I am sure, Flanagan's Principle of Minimal Psychological Realism. So I assume that acts of virtue, as I have defined the term, may require that the agent performing it exercise some self-control.

The final point I wish to deny is a strong version of the thesis of the unity of the virtues: the idea that one cannot, or cannot *really*, have any one of the virtues without having all of them. In fact, I think it very likely that one's efforts to develop one's powers of self-control with respect to actions of one type will often fail to transfer over to actions of another type, and so one can certainly be said to have one virtue but not to have another. We will see below, however, that it is not possible to possess the virtue of *phronesis*, practical wisdom, without possessing at least all the other core virtues. This weaker version of the unity thesis may be the one that Aristotle actually endorses, and so it may be a piece of the baggage of post-Aristotelian virtue ethics that I wish to avoid (*NE* VI.13 1144b30–1145a6; *Eudemian Ethics* VIII.1 1246b4–36; Reeve 2013, p. 261).

With our interlude complete, we can return to the main argument. My present aim is to defend my claim that the important interests of others may ground imperfect duties, and I will now present that defense, making essential use of the notion of virtue I have just finished elaborating. The crucial thought is that a plausible way of characterizing generosity is as the willingness to sacrifice the satisfaction of one's own peripheral interests (and the peripheral interests of those close to one) in favor of advancing the important interests of others. I do not have a principled argument to support this characterization, and I do not know what one would look like. But I do think many would join me in judging someone who conducted himself in this way as generous. And likewise, when one omits to advance the important interests of others so that one may make a definite contribution to one's own central ends, or

[7] This effort, and the inner conflict that necessitates it, is only lacking in the *phronimos*. As I discuss below, the *phronimos* is the *authentic* agent *par excellence*: his affective attachments are always in line with his judgments. Of course, the true *phronimos* also lacks any bad ends—so we can preserve a distinction between him and the *enkratic* agent. *Phronesis* is an ideal to be aspired to.

the central ends of those close to one, this does not make one ungenerous. Such actions are entirely consistent with leading a life that is generous on the whole.

The satisfaction of our peripheral interests is, *ceteris paribus*, a good of some kind; I do not wish to deny this. But since generosity is a virtue, and a good and worthwhile life is at least in part a virtuous life, acting ungenerously impairs one's leading a good life. And of course, it also impairs someone else's leading a good life—the one with the important interest to whom one might be generous. So when we sacrifice generosity for the sake of satisfying our own peripheral interests, we are attempting to realize a good in a way that undermines one of the requirements for leading a good life. That violating this requirement undermines part of what makes a good life possible is a reason not to act on the reasons to satisfy one person's peripheral interests when doing so conflicts with this requirement. Acting for reasons of that type undermines that for the sake of which one would act for reasons of that type. This is itself a reason not to act for those reasons. The important interests of others, therefore, ground exclusionary reasons in addition to first-order reasons, and competing reasons to satisfy peripheral interests fall within the exclusionary scope. But if I am right that it does not make one ungenerous to prefer to make a definite contribution to one's own central ends, then the reasons to perform these acts lie outside the exclusionary scope. When faced with a choice between advancing one of my own central interests and advancing the important interest of someone else, I am not under a duty to do the latter even if, on the overall balance of reasons, it is what I ought to do. I do no wrong by choosing the former. The duties grounded by the important interests of others are thus imperfect.

The claim that we are only free to neglect the important interests of others in order to pursue those long-term projects and goals that give our own lives meaning (or that give meaning to the lives of our nearest and dearest) is a strong one, and is meant to be so. There are only so many such projects that one can incorporate into one's life, and only so much time, energy, and resources one can devote to them before diminishing returns set in. And there is only so much leisure that is necessary for one to maintain the energy to pursue those projects. The average middle-class (or better-off) inhabitant of the developed world undoubtedly devotes a good deal of time, energy and resources to the satisfaction of preferences that are not connected to his life's main projects and goals. We are naturally resistant to the idea that there is something morally wrong in this. But given the prevalence of suffering and deprivation that exists in the world, this is not a conclusion we should shy away from. The extent of our moral duties may be uncomfortably broad; but if so, we can only be morally responsible agents by facing up to this uncomfortable fact.

Not all important interests, however, are created equal, nor are all central ends equally central or all peripheral ones equally peripheral. The precise extent of the scope of an exclusionary reason grounded by a person's important interest will depend on many situation-specific factors. Just a few of these are (1) how significant a sacrifice would be required for the person to satisfy the interest himself; (2) whether or not it is an interest which would be better satisfied if he satisfied it himself, even if it could be satisfied by someone else; (3) to what extent the agent is at fault for not being able to satisfy the interest—it may be too harsh to say that the

interest cannot ground an exclusionary reason unless the agent is entirely blameless for this; (4) which type of important interest it is—we would expect some of these, such as interests in the bare necessities of leading any valuable life whatsoever, to ground reasons with wider exclusionary scope than other types of important interest; (5) the odds of someone else satisfying the important interest if one does not do so oneself; and (6) the possibility of satisfying the important interest at a later time if one does not do so presently.

The duties grounded by interests that are simply important are not just imperfect, but also undirected—imperfect duties that are not owed to any particular person or class. There is often insufficient ground for identifying one person rather than another as being particularly well situated to advance important interests. My ideal situations for donating to a worthy charity, for example, are often ideal situations for many others to do the same. And we often cannot say that everyone who is in such a position should advance that charity's interest, since either the charity's interest will not be sufficient to justify such a claim, or there will be many other equally worthy charities, and we cannot each support all of them. So there is normally insufficient reason for holding that the actions of *this* person are especially relevant to *this* charity's interest. In general, the undirected imperfect duty of charity is owed to the set of those who are in need. It lacks a specific target. One fulfills such a duty by discharging it with respect to some member or members of this set. The duty does not require that it be discharged with respect to all of them, since to do so while still advancing one's own interests would be impossible, and the duty does not exclude reasons to advance one's own interests. Since the duty normally fails to pick out any particular individual or group, one is free to choose whom one will benefit from one's charity from a large pool of possible recipients. This explains why we typically think of an imperfect duty as one that the bearer of the duty can satisfy in the way he chooses. The interests that ground the duty do not determine when, for whom, or precisely how the duty is to be fulfilled.

The fact that undirected imperfect duties advance important interests does give the other members of my community a reason to urge me to fulfill those duties. This is the way in which imperfect duties are imposed on us. This imposition is justified by the fact that the interests that ground these duties also ground reasons for attempting to ensure that the duty-bearer fulfills them. Since these interests pass the test of importance, we have reason to urge one another to fulfill these duties. By contrast, in cases of what someone ought to do but does not have any duty to do, we are justified only in doing something weaker, like recommending or advising. We are not justified in censuring or sanctioning those who violate imperfect duties. The interests that ground these duties do not ground reasons to punish those who violate them. But there is an appropriate negative consequence to violations of imperfect duty. We at least ought to feel guilty, since the violation is an act of doing something wrong.

3.2 Directed Imperfect Duties

The second feature in virtue of which an interest can ground a duty is relevance to a distinct person or class. When an interest has the feature of relevance, it normally grounds a stronger reason for a distinct person or class to advance it than it does for anyone outside that class. It also grounds an exclusionary reason. As with the feature of importance, we need some precise way of determining when an individual or class is relevant to satisfying an interest.

There are two main ways in which one can be interest-relevant. The first is by being singled out by an interest. A is singled out by B's interest if B has an interest not only in someone ϕ-ing, but in A ϕ-ing. For example, children have important interests in being cared for in all sorts of ways. But over and above these interests, they have interests in being cared for *by their parents in particular*. These interests single their parents out. A child whose parents fail to care for him may still be well cared for, but nonetheless he has interests that have not been met. Being singled out in this way, parents have a stronger reason to care for their children in particular than they do to contribute to the care of others. In this case the interests ground agent-relative as well as agent-neutral reasons; some of the reasons grounded by the interests cannot be specified without reference to the agent for whom they are reasons.

An interest need not ground an agent-relative reason for there to be some distinct person who is relevant to its satisfaction. The second way in which one can be relevant to satisfying an interest is by being especially well-situated to satisfy that interest. To be especially well-situated is to be either necessary to the satisfaction of an interest, or to be significantly more likely to succeed in satisfying it (or in satisfying it to a greater degree) than anyone else, on account of one's resources, abilities, or proximity to the interest-holder (rather than on account of being singled out by the interest). Suppose someone were to faint at a party at which only one guest was a medical doctor. The person who fainted has an interest in being looked after, and this interest grounds an agent-neutral reason for someone to look after him. Everyone at the party has this reason to do so. But the doctor, in virtue of his special training, has a much stronger reason than anyone else to be the one who looks after him; only he has a reason to insist that he be the one who looks after him. Here, there is a distinct person who is relevant to the interest's satisfaction, even though the interest does not ground an agent-relative reason.

When an agent is relevant to the satisfaction of another person's interest, that interest also grounds an exclusionary reason for the relevant agent. The argument for this claim is similar in structure to the argument I used above to set the exclusionary scope of undirected imperfect duties. When we satisfy interests to which we are relevant, we often do so by performing acts of generosity. But other virtues also come into play in these cases. On those occasions where doing the good one is especially well situated to do requires exposing oneself to danger, the action one is called to perform expresses courage. More generally, though, these actions are expressions of the virtue of friendship, in one or another of its many forms. By

"friendship," I mean not only the sort of relationship which we usually denote by this word, but also familial relationships, and what Aristotle called "civic friendship": the relationships of mutual aid that hold together a political community. To these I would add what we might call "institutional friendship": the relationships of mutual aid that exist between co-workers at a well-functioning institution, which are closer than civic friendships but not so close as proper friendships. It is the interests of our friends, in all of these senses, which single us out, or which we are normally well-suited to advance. The closer the variety of friendship, the more interests of the friend to which we will be relevant, and so the more sacrifices will be morally required from us. Friendship, like generosity or courage, is an important part of a good and worthwhile life, and so fulfilling its demands to some extent is a genuine moral requirement. There are duties of friendship, and these duties partly constitute the relationship of friendship.[8]

It is likely that a full taxonomy of directed imperfect duties would break this class up into many varieties. The agent's particular circumstances will determine, first, what sort of friendship he is in a position to express; and second, what competing reasons he could act on instead, without thereby failing to exhibit the virtue of the relevant sort of friendship. The agent's circumstances will thus determine the exclusionary scope of the reason to advance the other person's interest, and thus determine the precise strength of the duty to do so. Some factors that will likely be significant here are: (1) whether or not the interest actually singles some person out; (2) how many individuals are well-situated to advance an interest that does not single any one person out, and how well-situated they are relative to one another; (3) the closeness of the relationship between the agent and the interest-holder relative to the closeness of the relationship between the agent and any other person or persons whose interests the agent would have to neglect—an agent may be relevant to the competing interests of two individuals but more relevant to one; (4) the centrality of the important interest or interests which the agent has an opportunity to advance relative to those he would have to neglect (centrality being to some extent a matter of degree); (5) whether or not the agent will have another opportunity to advance the interests which he would have to neglect on this occasion; and (6) whether or not anyone else is likely to advance important interests which the agent would have to neglect on this occasion. Depending on these factors, one person's relevant interests will ground exclusionary reasons with a range of scopes. The narrower scopes might fail to capture reasons for the relevant agent to advance even some of his own peripheral ends (since peripherality, like centrality, is a matter of degree). The wider scopes would capture much more than this, and so require the agent to sacrifice more. These factors will also help determine the overall balance of first-order reasons, and so determine whether, in the first place, the agent should sacrifice something important to himself, or whether he should omit helping someone with a very close relationship to him for the sake of helping someone with a more distant relationship.

[8] For a discussion of duties of friendship, and how they differ from duties of disinterested generosity, see (Raz 1994, pp. 29–43).

Since directed imperfect duties are genuine duties, there is an appropriate negative consequence for violating them. In addition to feelings of guilt for having committed a wrong, feelings of shame are also appropriate.[9] In the case of undirected duties, there was no particular person or class in whose eyes one could see oneself as having failed. But when I violate a directed duty, it is appropriate that I see myself as having let others down in their own eyes, as having failed to live up to a standard that they rightfully applied to me. Others are also justified in urging me to fulfill the duty, and if I violate it, the one to whom it was owed is justified in expressing disappointment, in making me feel some degree of shame for my wrong, and in criticizing my behavior.

3.3 Perfect Duties

According to Raz, an interest which is both important and relevant will ground a perfect duty if it also possesses a third feature: its satisfaction contributes to the common good. Interests with this feature do ground stronger first-order reasons than interests without it. They will succeed in grounding conclusive protected reasons in a wider range of cases then interests whose satisfaction is only a private good. That satisfying an interest merely contributes to the common good, however, is not enough for that interest to ground a perfect duty. Perfect duties are those duties that have the widest possible exclusionary scope, and so failing to satisfy the type of interests that ground them is practically always wrong. The interests that ground perfect duties are the ones whose satisfaction is necessary to the flourishing of the society in which the interest-holder lives. These are the duties that each person must fulfill for every other person in that society, in order to maintain a good and well-ordered communal existence. The interests that ground them are shared by all members of that society, since they are those interests which must be satisfied for a good social life to be possible. The exclusionary reasons grounded by these interests have the widest-scope: all types of competing reasons are normally excluded by them. The exception is when such a reason conflicts with another reason grounded on an interest with all three of these features. In such a case, there will be no one action one ought to perform, all-things-considered, and neither reason will exclude the other. The agent caught in the conflict will be morally free to act on either. Once I have established these claims, I will argue that despite grounding perfect duties, interests with only these three features fail to ground moral rights.

Here is a (fairly Aristotelian) argument that interests ground exclusionary reasons when their satisfaction is required for society to flourish. That the interests in this class be satisfied is a requirement for a society to function well, for the collective life of that society to be a healthy one. For (almost) all of us, part of leading a good life is leading a social life. If society is not functioning well—due to

[9] Walter Sinott-Armstrong also suggests that shame is appropriate when an imperfect duty is violated (Sinott-Armstrong 2005).

wide-spread violence, disorder, discrimination, poverty, etc.—it will be impossible for many, if not all, of the members of that society to lead healthy social lives. Since this is part of leading a good life, living in a well-functioning society is a condition of possibility for leading a good life. It may be possible in a particular instance to advance one person's interest by violating another's even when the latter interest is of the sort that must in general be satisfied for society to flourish. But to violate an interest of this type is to undermine the healthy functioning of society, and thus to undermine part of what makes any good (non-solitary) life possible. That violating these requirements undermines what makes a good life possible is a reason not to act on the reasons to advance one individual's well-being when doing so conflicts with these requirements. Acting for reasons of that type undermines that for the sake of which one would act for reasons of that type. This is itself a reason not to act for those reasons. Therefore, interests whose satisfaction is necessary to a flourishing society, in virtue of that necessity, ground not only reasons for others to advance them, but also reasons not to act for competing reasons, even when the competing reasons are grounded on other genuine interests. These interests, moreover, also ground reasons for the rest of society to actively discourage their violation, and to blame and censure those who do violate them. The obvious candidate for the virtue one exercises in respecting one's perfect duties is the virtue of justice—giving to each what is owed to him.

3.4 Moral Duty and Particularism

We must be careful to note that the account of moral duty I have been developing is consistent with my commitment to moral particularism. I do not assume the existence of any principles such as "*Ceteris paribus*, one should act so as to advance the important interests of others." It is perfectly true that I have claimed that the important interests of others normally ground reasons which outweigh those grounded by one's own peripheral interests. But this is not an assertion of a principle which serves to *explain* the rightness of an action that advances an important interest rather than a peripheral one. Rather, it is an observation of an abstract ethical regularity. What explains the fact, in a particular case, that one action is right, is that the reasons favoring that action, taken together, are stronger than the reasons against it and in favor of another. And what explain the fact, in a particular case, that one reason outweighs another, are the features of that particular case—features like the fact that one agent will be unable to continue his pursuit of one of his central ends unless some action is performed. Again, I do claim that an interest which must be satisfied, in order for someone else's continued pursuit of a central end to be possible, normally grounds a stronger reason than does an interest of one's own that can go unsatisfied without such severe consequences. But this is just an observation of the same regularity, expressed in terms of one of the features which I include under the heading of 'importance.' There is no explanatory work to be done that is not done by the individual reasons present in particular cases and the features of those cases

that affect the weights of those reasons. We will examine the process whereby the features of a particular case determine the weights of the reasons present in that case in greater detail, when I introduce my model of particularist moral reasoning at the end of this chapter.

The account of exclusionary reasons is also meant to be a particularist one. Let us first define holism with respect to exclusionary reasons as the claim that interests do not ground exclusionary reasons with a fixed scope. The particularist will then assert that healthy moral reasoning requires no principles such as "An interest in pursuing/achieving end X grounds an exclusionary reason which captures the first-order reason grounded by an interest in pursuing/achieving end Y" or "*Ceteris paribus*, disregard the reasons grounded by one's peripheral interests when one can advance another's important interest." To claim that another's important interest, for example, grounds an exclusionary reason that captures the reason grounded by one's own peripheral interest is once again to observe an abstract ethical regularity. It is not to assert any principle with an explanatory role in the account. The scope of an exclusionary reason in a particular case is determined and explained by the extent of the personal sacrifice which is consistent with acting virtuously in that case. And just how much sacrifice is required is not determined by any principle, but by the sum total of the many morally relevant features of each case (just a few of which I have attempted to enumerate).

The sorts of regularities I have been discussing would not even make good models for heuristic principles in moral reasoning. Interests in specific ends are not in and of themselves central or peripheral, important or relevant. The nature of the interests at play in a particular case depends on the specific facts of the lives being led by the agents involved at that time, and of the society in which they live. And for a given agent at a given time, not all central interests are equally central, nor all peripheral interests equally peripheral. Likewise, the sorts of interests I have grouped together under the headings "important" and "relevant" are not all created equal. I have already discussed a few of the many circumstantial factors that help to determine the strength of the first-order reason and the scope of the exclusionary reason grounded by such interests in particular cases. And so even to use a principle like "*Ceteris paribus*, act so as to advance the important interests of others" as a heuristic would first require a fairly extensive examination into the details of the particular case to determine which interests are important and which are not, and whether *ceterae* really are *pares*. The deliberator who requires the expedience of a heuristic is better off using a virtue-based rule, like "Be generous," bearing in mind that it is merely a heuristic.

The point of categorizing interests and duties as I have done, and of eliciting the patterns and regularities that seem to exist among the interests so categorized and the reasons they ground, has not been to derive a set of moral principles. We have seen both that there is no explanatory work for such principles to do, and that, since the regularities that exist are crucially dependent on such a wide array of context-specific factors, the usefulness of heuristic principles based on these regularities for moral reasoning is doubtful. Rather, the point is to lend some structure, some organization, to the account of moral duty. There are as many different types of moral

duty as there are pre-emptive reasons—as there are weights of first-order reasons and scopes of exclusionary reasons. These weights and scopes are determined and explained in each case by the combination of the members of some subset of all the many different types of morally relevant features, some of the most common of which I have just discussed. This method of categorizing moral duties—perfect vs. imperfect, directed vs. undirected—and the features in virtue of which interests ground duties—importance, relevance, necessity for a given society to flourish—is meant to be a rough guide to the moral jungle, a way of representing the complex realm of moral duty using a much tidier system of common ethical concepts. We are not trying to carve the normative realm at its joints. These categories simply serve the purpose of providing a framework in which making the argument that human interests can ground moral duties becomes a manageable task. These regularities, moreover, are of considerable interest in themselves. We are trying to discover what we can about the nature of moral duty. To identify the types of interests that normally ground duties, and determine the shared features in virtue of which they are typically capable of doing this, adds significantly to our knowledge and understanding.

This concludes my account of moral duty, and my answer to Anscombe's challenge. Human interests do ground pre-emptive reasons with a variety of strengths and scopes. I will now argue that interests possessing only the three features discussed thus far *fail* to ground rights *despite* grounding perfect duties, and thus that the grounding of a duty is an intermediate step in an interest-based account of the existence of rights.

3.5 Exactable Duties and Rights

Coercion is not necessarily justified whenever a duty is violated. As Mill rightly observed, some of them should be secured not through coercive means but through the effects of "the opinions of [our] fellow creatures" (Mill 1863/2002, p. 48). Duties grounded on interests with these three features should at least be secured through non-coercive social pressure. But not all of them should be coercively exacted. It is possible that a society would fare even worse if all such duties were exacted than it would be if it tolerated occasional failures to fulfill some of them, and sought to minimize these failures through customary norms. This is the first flaw in Raz's theory. Interests with only the first three features do not ground a sufficient reason for coercion when they are violated, and so possessing these three features is insufficient for an interest to ground a right. This is why I add a fourth feature to Raz's list. In some cases, coercing those who fail to do their duty has a negative social impact that outweighs the benefit, to both the interest-holder and the public, of securing the duty. For an interest to ground a justifiably enforceable duty, this must not be the case.[10]

[10]The order in which the features are given is important. We must determine whether an interest possesses the third before inquiring about the fourth. Suppose we have determined that an interest

Consider an example. A democratic government has a moral duty to permit each of its citizens to vote, and to protect their exercise of that liberty, regardless of their sex, race, or religion. Suppose such a government fails to do this. We want to know whether this duty can be justifiably exacted from the government. If it can be, the disenfranchised individuals, along with those who support them, will have to exact it from the state themselves. Civil disobedience may not be enough to accomplish the task; some coercive measures may be necessary. There will thus be a cost associated with exacting the duty. Some level of social disruption will follow. In a case like this, however, the duty is enforceable. The cost in social disruption is outweighed by the harm of continuing to disenfranchise a segment of the population. The situation is intolerable. The exactability of the duty is of course qualified; the harm caused by continued disenfranchisement does not outweigh all possible coercive means to exact the duty, however extreme. But the duty is enforceable nonetheless.

Contrast this duty with the duty we each have to show a minimal level of respect for the differing cultures, ethnicities, and religions of those around us. A society in which this duty is not fulfilled is not a healthy one. But to live in a society that punishes everyone whose actions express prejudice is worse than to live in one that tolerates such views up to a point and attempts to minimize them by non-coercive means. I do not mean to deny that there is some level of respect that is enforceable. I am only claiming that it is possible to fail in one's duty of respect without there being sufficient reason to exact the duty. This is so even though I ground duties on interests whose satisfaction is essential to a flourishing society. A community can make some group within it feel so unwelcome—*without* making them feel that they are ever in any danger—that all members of the group gradually move somewhere else. Their actions make it impossible for all the members of that community to live together, by denying some members entrance into the life of the community. They fail to fulfill an important duty and thus do a great wrong. It is entirely appropriate that they be discouraged from, and censured for, doing so. But it is doubtful whether it would be possible to for a society to use coercion to discourage and correct such behavior without thereby creating an atmosphere in which its members fear persecution for expressing disapproval of others, which can be perfectly legitimate. In this case, then, coercion is unjustified even though an interest with the first three features is violated. The value of satisfying such interests cannot, as Raz claims, be sufficient to ground rights.

An interest with all four of these features not only grounds a duty, it grounds a duty whose fulfillment can be justifiably coerced.[11] The justification, moreover, sat-

---

is both important and relevant, and then ask whether the good that would come of coercively obtaining its satisfaction outweighs the bad. If we have not first determined whether its satisfaction is necessary to the flourishing of society, we will end up including in our judgment the harm to the coerced of depriving him of his freedom. If he is in fact not morally free in this case, we will give too much weight to the side that disfavors coercion.

[11] Just how significant and central a contribution is made, or how much harm is caused by failing to fulfill the duty, as well as how urgently the duty must be fulfilled, will set limits on what sort of coercion is appropriate and on who can apply coercive measures.

isfies our constraint. The bearer of the duty is not morally free to refrain from satisfying the interest. The interest grounds an exclusionary reason for him to disregard any reasons that compete with his reason to satisfy the interest. So there is no valid objection to coercion based on his interest in being free to determine his own actions. In virtue of possessing the fourth feature, the interest grounds a reason for coercion which outweighs any remaining reasons against using coercion. In grounding enforceable duties on interests, I have nowhere assumed that the interest-holders possess any rights to have their interests satisfied. By the correlativity of rights and duties, for one to be owed a duty whose fulfillment can justifiably be secured with coercion implies that one has a right against the duty-bearer. It follows that the existence of rights can be explained in terms of interest-holders being owed duties grounded on interests with these four features. The ground of these duties also grounds the right sort of justifying reason to exact the duties, and so enforcing their correlative rights is likewise justified; the first justification is tantamount to the second.

We can now see why the second flaw in Raz's theory, which I discussed in the last chapter, is a deep structural one present in any teleological theory with a structure similar to his. Recall that Raz makes three claims: (1) rights are grounded on the value of satisfying the interests they protect; (2) when duties are correlative with rights, the duties are based on the rights; and (3) duties are pre-emptive reasons. There is good reason to accept the third claim, and I do accept it. To satisfy the justification constraint, the first claim would have to be modified, so that a right would only be grounded if the interest also grounded an exclusionary reason to disregard any reasons against satisfying it. In order to preserve claim (2), the Razian would then have to argue that a conclusive first-order reason to advance the interest is not grounded until the right is grounded. But I have shown this claim to be false. An interest that meets Raz's criteria (of important, relevance, and contribution to the common good), and grounds the required exclusionary reason (by being *necessary* to the common good), thereby grounds both a first-order reason and an exclusionary reason of sufficient scope to normally capture all potential competing reasons. That is, the interest grounds a perfect duty. An interest with only the first three features, furthermore, fails to ground a right even though it succeeds in grounding a perfect duty. A right does not come into existence until a justifiably enforceable duty is grounded by an interest possessing all four features. So we cannot explain the existence of rights by appealing to the interests rights protect without establishing, as an intermediate step, the fact that those interests ground perfect duties. My theory reverses the usual order of explanation between rights and duties, and succeeds in explaining the existence of justifiably enforceable duties—and thus in explaining the existence of rights—in a way that satisfies the justification constraint.

3.6 Defending the Direction of Explanation

In addition to justifying the enforcement of rights, there is another advantage my theory has over one like Raz's, which attempts to ground rights on the value of satisfying interests. Raz and I agree that rights are not absolute. In my case, this is because an interest that grounds an enforceable duty in one set of circumstances may fail to do so under other circumstances in which the interest does not possess the fourth feature. For Raz, the interests that normally ground a right may be defeated by stronger opposing considerations (Raz 1986, pp. 183–184). When a right is defeated, the force of the interests on which the right is grounded are weakened or overcome; when this happens, "no one could justifiably be held to be *obligated* on account of those interests" (Raz 1986, p. 184). Duties collapse along with the rights that justify them. This is a consequence of claiming that duties exist in virtue of rights. Though Raz acknowledges that not all duties are based on rights, those that are—the duties to advance individuals' interests—must fall with the rights on which they rest. If conflicting considerations lead us to deny that someone has a right to something, because his interests cannot support a right, we must also deny that those interests ground a duty to her. If we wanted to maintain the existence of a duty, we would have to find some other way to ground it. This is an undesirable result. We should be able to acknowledge that someone's right is defeated in a particular case without denying the force of the interests on which that right usually rests and without denying that those interests are still capable of grounding a duty. Since my theory claims that rights exist in virtue of duties, it has the happy consequence that a claim to a right can be defeated without eliminating the corresponding duty. We can recognize what we owe to each other even when it is not for the best to recognize a right to what is owed.

Raz has two arguments for the claim that when a duty corresponds to a right, the duty exists in virtue of the right. Neither argument undermines the view I am offering, so I will only address them briefly. The first is that "one may know of the existence of a right and of the reasons for it without knowing who is bound by duties based on it or what precisely are those duties" (Raz 1986, p. 184). Raz gives the example of a child's right to education: we may know that children have a right to education without knowing what sort of education or who has a duty to provide it. Raz claims that even if knowledge of the content of a right is incomplete, it still counts as knowledge that someone has a particular right (Raz 1986, p. 185). One cannot know of a duty, however, without knowing who bears it and to whom it is owed. I do not wish to dispute any of this. This sort of epistemic priority is not what I am interested in. The fact that rights have this sort of priority over duties does not show that they have explanatory priority, which is what concerns me. There is no inconsistency in claiming both that we have rights in virtue of being owed duties, and that one may learn of a right's existence without also learning who bears the duty correlative with that right. Raz goes on to argue that "If a duty is based on a right, on the other hand, then it trivially follows that one cannot know the reason for it without knowing of the right (or without knowing that the interest which it pro-

tects is sufficient to be the ground of a duty—which is the definition of a right)" (Raz 1986, p. 185). I do not deny that one cannot know the reason for a duty without knowing that the duty protects an interest that is sufficient to ground it. Instead, I am questioning the claim that rights are grounded on the value of satisfying the interests they protect.

The second argument is based on the fact that rights are dynamic: what it means to have a right to education, for example, may change over time (Raz 1986, p. 185). Raz makes the further claim that "With changing circumstances, [rights] can generate new duties" (Raz 1986, p. 186). But we need not infer from the dynamic character of rights that rights give rise to duties. It is not because children possess a right to education that new duties toward them arise with changing circumstances. Rather, with changes in circumstances come changes in interests, and in whether a given interest possesses the features it needs to ground a duty. Raz is correct that in a sense a right, such as the right to education, may persist through these changes, even if the duties that correspond to it change. This simply indicates that the duties grounded on the changing interests continue to give rise to what is at some level of generality a right to education. The precise content of the right, however, changes with the duties on which it rests. Changes in interests can also alter duties in such a way that new rights emerge, or that old rights are extinguished. Interests, duties and rights are all dynamic. This dynamic quality does not, however, support an explanatory priority of rights over duties.

4 Neo-eudaimonism Part V: Deliberation, Ethical and Otherwise

A good life is, among other things, a life of excellent deliberation about both ends and means. But it is also a life of excellent *ethical deliberation*, of reasoning about what one ought to do in particular circumstances. We are finally in a position to complete our account of the deliberative aspect of a good life by discussing ethical deliberation, and its relation both to ends-deliberation and to the theory of duties and rights just developed.

4.1 A Closer Look at the Virtue of Phronesis

I have already characterized *phronesis* as excellence in ends- and means-deliberation. I now expand that characterization to include excellence in ethical deliberation. My first task is to discuss ethical deliberation in itself, and to describe excellence in this type of deliberation. I will come later to the question of how ethical deliberation and ends-deliberation relate to one another, and in particular to the possibility of conflict between them. In exploring how that conflict is resolved, we will see why we should

associate ethical deliberation with the same capacity and the same virtue as ends-deliberation.

An agent can act *for* a reason without being *aware* of the fact that that is the reason for which he acts. It would be a baseless over-intellectualization of human experience to claim that whenever we are moved to act by our recognition of some fact which does favor our performing that action, we must be aware that the fact so favors the action and that we are moved to act by it. This need not even be the case if, later on, having been asked what our reason for acting was, we are on reflection able to identify the reason for which we acted. In some cases, however, we are aware of all this both before and during our action. In such cases we act deliberately. Ethical deliberation is the process, preceding action, by which we identify the reasons that bear on us in a given situation, determine their valences and weights, and arrive at a conclusion about which action is most supported—and thus, a conclusion about which are the reasons we should act *for*. This much should be familiar ground. Excellence in this activity is then easy to characterize, at least on a very general level. It is excellence in identifying the reasons that bear on one's action, appropriately weighing them, and reaching a conclusion about what to do on that basis. This is *phronesis* with respect to ethical deliberation. What exactly it means to do these things excellently is a question I address in greater detail below. I would venture in addition that identifying (non-deliberatively) whether one is in a situation in which ethical deliberation is appropriate (as opposed to one in which it is more appropriate that one act on a hunch or an instinct, or one in which action is needed with an urgency that precludes deliberation) is itself another aspect of exercising *phronesis*.

We will say that an agent *has* the virtue of *phronesis* with respect to the aspect of ethical deliberation when the agent (a) regularly and reliably engages in excellent ethical deliberation and acts on the conclusions of that deliberation; and (b) has well-developed self-control with respect to engaging in excellent ethical deliberation and acting on the conclusions of that deliberation in the face of disabling or interfering environmental factors. This is why *phronesis* is the "crown of the virtues." One cannot exercise it without acting generously, courageously, justly, etc. We will say that an agent has the virtue of *phronesis* with respect to its aspect of means- and ends-deliberation when the agent (a) regularly and reliably engages in excellent means- and ends-deliberation and acts on the conclusions of that deliberation; (b) achieves authenticity by training his emotional attachments, through a process of habituation, to track his value judgments (this is just an Aristotelian interpretation of Bradley's mechanism of cultivating a taste for reasons, which I discussed in Chap. 4—it reflects the Aristotelian idea that the practically wise agent has the right moral emotions); and (c) has well-developed self-control with respect engaging in excellent ends-deliberation and acting on the conclusions of that deliberation in the face of disabling or interfering environmental factors. Of course, the *phronimos* must possess *phronesis* with respect to all its aspects simultaneously. This means that all forms of deliberation must be integrated into a single coherent process of practical reasoning. That is the task for the end of this chapter.

4.2 Particularism Meets Bayesianism

Let us now take a closer look at what it means to engage in excellent ethical deliberation, from a particularist perspective.[12] Recall Dancy's definition of reasons-holism:

1. What is a reason in one situation may alter or lose its polarity in another.
2. The way in which the reasons here present combine with each other is not necessarily determinable in any simply additive way. (Dancy 2003, p. 132)

I urged at the end of the last chapter that we *not* interpret (2) as entailing that the normative support given to a particular action *in a particular situation* cannot be modeled as the sum of the numeric representations of the weights of the reasons that favor it *in that particular situation*. Rather, it should be interpreted as the claim that, just as reasons have no fixed valence (for what is a reason to do something in one situation may be a reason against it, or no reason at all, in another), so too they have no fixed weight. The extent to which a reason counts in favor (or against) a given action may differ from one situation to another. There is no pre-set weight which any reason carries with respect to any action in every situation in which it is a reason at all. Thus, in advance of looking at the details of a whole situation, it cannot be known (by looking it up in a "Table of Weights of Reasons") what weight any single reason will carry. This, I argue, should be our interpretation of Dancy's denial of "simple additivity." Whether this is what Dancy himself meant, I cannot be certain. But it is the only interpretation which seems to me to be defensible, and it is essential for the representation of particularist reasoning I will now present.

Dancy's particularist theory of moral reasoning employs five basic notions: that of a contributory reason, an enabler, a disabler, an intensifier, and an attenuator. The latter four are types of facts which are not themselves reasons, but which affect the valence and weight of reasons. An enabler makes it the case that a fact which would not be a reason for an action in the absence of the enabler is one. A disabler does the opposite. An intensifier makes it the case that the weight of a reason for a given action is greater than it would be in the absence of the intensifier. An attenuator does the opposite. The particularist account of the ground of moral duty given above is structured by the concepts which are employed in particularist moral reasoning. Many of the sorts of facts about the interests which ground reasons that have been discussed above—like the fact that an interest is central to an agent's life-plan, or that an interest must be satisfied in order for an agent to continue pursuing some larger goal—I take to be examples of intensifiers, at least in most circumstances: they normally make the reasons for others to act which are grounded by those interests stronger. But there are cases in which, for a given agent these facts are attenuators—the fact of an interest's centrality may make it inappropriate for one person to advance it, as doing so is another's responsibility or privilege. The fact that someone is at fault for his inability to satisfy one of his interests may be an example of a

[12] For an argument that Aristotle himself was a particularist, and thus that *phronetic* reasoning, even in classical Aristotelian eudaimonism, is particularist reasoning, see (Leibowitz 2013).

disabler, or an attenuator, or be normatively neutral, depending on the other circumstances of the case. Those who are unfamiliar with the basics of Dancy's theory should consult his work for a more detailed description and development of these notions, including many rich and illustrative examples. My purpose here is to provide a precise, formal characterization of these notions and of the particularist moral reasoning that employs them, using a Bayesian framework.

Suppose an agent encounters a choice situation and recognizes that his available actions are ϕ and ψ. He begins to consider what considerations bear on his choice between these. Take the term "considerations" to cover reasons, enablers/disablers, and intensifiers/attenuators. Assume that the agent is certain about the considerations he recognizes. So, for example, it will always be the case that, if the agent recognizes a reason R, he is certain that R obtains: $p(R) = 1$. This assumption is not necessary; I make it merely to simplify what follows (notationally, as much as in any other respect). Interpret $p(\phi)$ as the agent's degree of confidence that ϕ is the action he ought to perform. We assume the agent enters the situations with prior probabilities for $p(\phi)$ and $p(\psi)$, which reflect his views on each of these actions normally being the thing to do when the other is an option.[13]

Now suppose the agent observes some fact R_1 which he takes to be a reason for ϕ-ing.[14] If we assumed the agent were certain about the extent to which R_1 favors ϕ-ing, we would then write: $p_{new}(\phi) = p(\phi|R_1) = n > p_{old}(\phi)$, for some number n (recall our assumption that $p(R_1) = 1$). Recognizing R_1 would have straightaway made the agent more confident that ϕ is the thing to do. But we should not at this point assume that the agent has reached a conclusion about this. The number n is the agent's current *expectation* of $p(\phi|R_1)$: $exp(p(\phi|R_1)) = \int_0^1 p\left(p\left(\varphi \mid R_1\right) = N\right)\left(N\right) dN = n$. But the agent is open to updating this expectation before he begins updating his judgment of $p(\phi)$.

His next step is to consider whether there are any disablers which prevent R_1 from functioning as a reason in the present context. If he observes a disabler D_1, he

[13] These prior probabilities, along with the priors for each other type of hypothesis discussed below, reflect the agent's attempt to abstract regularities from his past experiences. The observation of regularities thus plays a role in ethical deliberation, even though principles do not.

[14] This notion of recognizing a reason is analogous to Aristotle's notion of deliberative appearance. The main difference between the two is that, since Aristotle conceives of ethical deliberation as syllogistic and as terminating in a general conclusion, deliberative appearance follows practical reasoning, and allows the agent to discriminate whether the situation he is currently in is an appropriate one for enacting a prior decision to perform a type of action. For example, if an agent has decided to donate to efficiently run charities, his deliberative appearance is what enables him to discriminate whether the charity he is currently considering supporting is efficiently run. One's deliberative appearances continue to sharpen throughout one's life, and this is why Aristotle sees *phronesis* as developing throughout one's life, even for the one who already deliberates as excellently as possible (Reeve 2013, p. 210). For the particularist, on the other hand, ethical deliberation has an inductive, rather than a deductive, structure (like ends-deliberation), and begins with the recognition of reasons, enablers, attenuators, etc.—reasons such as the fact that this organization is a charity, and attenuators such as the fact that it is run inefficiently.

will conclude that R_1 is not a reason in this context. Let $p_{old}(\phi) = m$. The agent will conclude that $p(\phi|R_1) = m = p_{old}(\phi)$. That is, he will conclude that R_1 lends no support for ϕ–ing. He reaches this conclusion by updating his expectation: $exp_{new}(p(\phi|R_1)) = exp((\phi|R_1)|\ D_1) = m$. The agent will then consider whether there are any other facts which he should recognize as reasons owing to the presence of enablers. Suppose that for some fact F_2, the agent initially judges $p(\phi|F_2) = m = p_{old}(\phi)$. But upon observing enabler E_1, he updates his expectation to $exp((\phi|F_2)|\ E_1) = n > m$. So the agent revises his provisional conclusion and now judges that $p(\phi|F_2) = m > n = p_{old}(\phi)$. That is to say, he now (provisionally) judges that F_2 is a reason to ϕ. We can now cash out the holist's claim that reasons have no fixed valence in terms of disablers and enablers. A fact which is normally a reason to ϕ may be disabled from playing that role by one feature of the present circumstances, and enabled as a reason *against* ϕ-ing by another feature.

So let us suppose that the agent has identified the reasons present. He will then determine whether there are any intensifiers or attenuators affecting these reasons. If there are, he will revise his provisionary conclusion about the extent to which R_1 favors ϕ-ing. Let us look at the case of an intensifier. Say that initially, $p(\phi|R_1) = n > p_{old}(\phi)$. We represent the result of recognizing an intensifier with another expectation update: $exp_{new}(p(\phi|R_2)) = exp(p(\phi|R_2)|\ I_1) = q > n$. The agent now judges that $p(\phi|R_2) = q$. Recognizing an attenuator works in a similar (though obviously opposing) way. The claim that a given fact is an enabler/disabler/intensifier/attenuator is itself a hypothesis, and the agent may at first provisionally accept such a hypothesis, but then go on to reject it. For example, an agent may initially suppose that the fact that another's interest is important intensifies his reason to advance it, but on learning that someone else is both responsible for and capable of doing so, reject the initial hypothesis and take the fact to be an attenuator. Finally, the agent may find that he has evidence for revising his hypothesis regarding the extent to which an intensifier or attenuator strengthens or diminishes a reason in a particular case. Each of these cases can be represented by appropriate updates of the agent's expectations. The agent will go through this process for each reason that he recognizes.

This process concludes when all considerations have been accounted for. The agent will then updates his $p(\phi)$ and $p(\psi)$ on the reasons he has recognized and arrive at final judgments regarding his available actions: $p_{final}(\phi)$ and $p_{final}(\psi)$. If $p_{final}(\phi)$, say, is higher than $p_{final}(\psi)$, the agent will judge that he ought to ϕ. These final judgments will be based on the agent's final conclusions regarding his hypotheses about the extent of the support given by each reason to the action it favors. That is to say, he will have judgments of the form $p_{final}(\phi|R_i)$ for each reason, just before he updates his $p(\phi)$ on those reasons. The agent can then reach a final judgment, $p_{final}(\phi)$, by starting from $p_{original}(\phi)$ and updating on each reason. We must make one very important assumption about this entire process. We assume that the order of updating expectations of conditional probabilities on intensifiers and attenuators, and then of updating judgments of actions on those conditional probabilities, does

not matter. The agent's updating, in other words, is *path-independent*.[15] This ensures that the agent will reach the same conclusion about what action is right, regardless of the order in which he considers the facts. This seems a reasonable assumption to make about someone who is deliberating well. Being influenced by the order in which one happens to consider the facts of a choice situation is an example of irrational cognitive bias.[16] With each update of $p(\phi)$ on a reason R_i, we can take the difference between the new and the previous $p(\phi)$. This is the value of the weight W_i of the reason R_i with respect to ϕ in this situation, as judged by the particularist deliberator following a particular deliberative chain. As far as I know, this is the only precise definition of the weight of reasons for action offered, despite the importance of this notion in contemporary ethics.[17] It is also just what we should expect from such a definition. From a pre-theoretical standpoint, the weight of a reason should be a measure of the change in the agent's mind about the likelihood that the action for which it is a reason is right. With each update of $p(\phi)$, the value of $p(\phi|R_i)$ for each of the remaining reasons will of course change in accordance with Bayes' Theorem. So the weight of each reason will depend on the order in which the agent updates his judgment on those reasons. But given path-independence, this will not affect the value of $p_{final}(\phi)$. And this does *not* imply that the weights of reasons are relativized to the deliberating individual. Rather, these weights are relativized to the particular deliberative chain which the deliberator follows. Two agents with the same priors and the same conditional probabilities will not only reach the same $p_{final}(\phi)$; each will also recognize that the weight given to each reason by the other is the same as the weight which he would give to the reason, were he to follow the same deliberative chain as the other. And the difference between $p_{original}(\phi)$ and $p_{final}(\phi)$ will be the sum of the weights, preserving simple additivity. Alternatively, to avoid this relativity, we can take the weight of each R_i for or against ϕ as judged by the agent to be $W_{i,\phi} = p_{final}(\phi|R_i) - p_{original}(\phi)$, and the overall weight of the reasons for and against ϕ as judged by the agent to be $W_\phi = \sum_i [p_{final}(\varphi \mid R_i) - p_{original}(\varphi)]$, which will not be equal to $p_{final}(\phi)$, but which will *track* its value—i.e. $p_{final}(\phi) > p_{final}(\psi)$ iff $W_\phi > W_\psi$ (still assuming for each R_i $p(R_i) = 1$; otherwise we must use $W_\phi = \sum_i \left[\left(p_{final}(\varphi|R_i)\, p_{new}(R_i) + p_{final}(\varphi|-R_i)\, p_{new}(-R_i)\right) - p_{original}(\varphi)\right]$).[18]

[15] For a statement and discussion of a path-independence condition that can be used in an axiomatic characterization of Bayes' rule, see (Majumdar 2004, p. 264).

[16] Specifically, such influence is likely to be due to framing effects and recency bias.

[17] Recall from Chap. 3 that Dietrich and List define a normative reason-weighting function, but have nothing to say about where these weights come from. We also saw that they fail to draw any connection between an agent's normative judgments and his actual actions. I develop an account of this connection below.

[18] Selim Berker has developed a very interesting argument to the effect that moral particularism is incoherent if it does not allow for the additivity of the weights of reasons. He assumes it does not, based on Dancy's definition of the view. But since I think that interpretation of Dancy is questionable, and since (more importantly) my characterization does allow for additivity, there is no point of contact between Berker's argument and my position. See S Berker (2007) "Particular Reasons" *Ethics* 118: 109–139.

The final normative judgments of an agent depend on his prior probabilities. Given my commitment to Aristotelian pragmatism, I am committed to the hope that any discrepancies between the judgments of different agents regarding the rightness of actions is temporary—that in the long run, these judgments will converge. We must be careful not to assume that because the deliberator has arrived at *final* normative judgments regarding his current situation, he is therefore *certain* about those judgments. Certainty in these matters could only be achieved in the same circumstances as convergence: *viz*., at the limit of normative inquiry. The agent must be open to revising his degree of confidence in any of his normative hypotheses regarding any particular case on the basis of discussion and debate with others, especially since some relevant evidence, such as the value another person places on one of his interests, must be communicated by that other person.[19] Ethical deliberation, as I noted in the previous chapter, is fundamentally social and dialogical. The impact of dialogical exchanges with others on a given agent's normative judgments can be modeled in the same way as the impact of testimonial evidence on a given agent's preference judgments is modeled in Chap. 4. The presence of a rule is an especially useful heuristic in cases in which different agents' views on what is right would otherwise initially conflict, and often alleviates the need for burdensome and time-consuming dialogue. But the question of whether a rule ought to be followed in a particular case is not necessarily dispelled just because we decide it is expedient to ignore it.

This concludes my presentation of particularist moral reasoning as represented in a Bayesian framework. Let us now consider the relationship between this type of reasoning and ends-deliberation.

4.3 Ethical Deliberation, Side Constraints, and Ends-Deliberation

The problem we must now face is: What happens when the result of ethical deliberation about a choice one is currently confronting conflicts with the result of one's ends- and means-deliberation? Such a conflict can occur even if we assume, as I will in what follows, that the agent has deliberated excellently over ends. One important type of evidence for the value of a potential end is the fact that one can normally pursue the end while respecting one's categorical reasons for action. But this does not exclude the possibility that an agent who has deliberated over his ends as well as anyone could will find himself in a situation in which the action with the greatest

[19] In fact, an agent's judgment of the weight of a reason—the difference which he appropriately takes the reason as making to his judgment of what he ought to do—may be doubly intersubjective, singly intersubjective and singly objective, or doubly objective. It will be objective in one sense if he is both certain and correct that the reason obtains, and objective in the other sense only at the end of inquiry, when $p(\phi|R)$ is given its ultimate value, from which it need never again be updated. A doubly objective judgment represents the reason's real weight.

expected value, based on his deliberative preference-ranking, is not the action which he judges, on the basis of ethical deliberation, to be the one most supported by the balance of reasons in this particular case. Solving this problem will require nothing less than a final reconciliation between the particularist approach to moral reasoning, and our decision-theoretic approach to the other forms of practical reasoning. We will see that these two modes of reasoning work in concert, in the practically wise agent's deliberations about what is to be done.

I will address this problem in two stages. First, I will consider the case of an agent who cannot pursue his end without violating his duty. Then I will consider the case of an agent who can pursue his end consistent with his duty, but cannot do so without acting against the balance of reasons—that is, the case of a conflict between one's own ends and the supererogatory pursuit of the interests of others.

The basis for my answer to the question of how to resolve conflicts between ends-deliberation and the recognition of duty, can be found in a remark of Sen's:

> [T]he violation of self-goal choice is arising here from the normative restraints we may voluntarily impose on ourselves on grounds of recognising other people's pursuits and goals, without in any substantive sense making them our own goals. (Sen 2007, pp. 353–354)

By "self-goal choice," Sen means choosing the action with the greatest expected value, based on one's preferences over freely adopted ends—in other words, acting in accordance with the conclusions of one's ends-deliberation. In Chap. 4 I argued that it is a normative and rational requirement that an agent act in this way. But there was an important point which I left out of that discussion (since we did not yet have the resources to introduce it). An agent must, whatever else he does, make a choice which he is free to make. One way in which an agent may be unfree to make a particular choice is by being physically incapable of making it. It may well be the case that there is some action which, *if* the agent were physically capable of performing it, would have a higher expected value in the circumstances than any other possible action. But if the agent is not so capable, we of course would not say that he acts irrationally, or does not act as he should, when he does something else instead. The point I want to make now is that an agent may likewise recognize that there is some action which, based on his deliberative preference-ranking, has a higher expected value than any other—but it is *not* an action which he is *morally free* to perform. He cannot perform it without violating a duty which he has in this particular case. Insofar as he respects his duty, the agent will treat that duty as a side-constraint. What that means from a decision-theoretic perspective is that he will revise his conception of his choice situation, so as to exclude those actions which he is not morally free to perform from the set of available actions. He will then choose the action with the highest expected value from within this revised set of available actions. And in doing so, the agent no more flaunts any normative or rational requirement than does the agent who limits himself to choice among those actions he is physically capable of successfully performing.

My expected value principle should be taken as presuming that the agent will be choosing only among actions he is free to perform—morally as well as in any other

relevant sense. Ethical deliberation about one's duty—that is, ethical deliberation that focuses on exclusionary reasons—must therefore *precede* instrumental deliberation. Before the agent even considers his preferences over the potential outcomes of his available actions and his judgments of the efficacy of those actions, he must determine whether there is any basis for eliminating any of those actions from consideration. He must shape the choice situation he is about to consider in accordance with his moral duty. Of course, from a *psychological* perspective, Sen is correct to call this restriction of one's choices voluntary: agents are psychologically capable of disregarding their duty. Recognizing one's duty—through a recognition of one's exclusionary reasons—and respecting that duty—by *acting on* those exclusionary reasons—which is what the agent does when he chooses to view his choice situation as restricted to the unexcluded options—are acts of *phronesis*. Even an agent who has developed the capability to perform these acts may fail to perform them on a given occasion, owing to insufficient self-control (which may have any number of causes).[20] To restrict the choice situations one allows oneself to consider is a form of moral strength; it is akin to the self-control shown by the *enkratic* agent, since it is a way of shielding oneself from the temptation to form intentions one ought not have (in this case, intentions to pursue ends, where the reasons for pursuing them in the circumstances are all excluded). But we must be careful to remember that this imposition on oneself of duties as side-constraints is not voluntary in another way. The agent is under the duties he is under whether he likes it or not; he has no choice in that matter. The agent who disregards his duty is ignoring a real and relevant feature of his choice situation, and ignoring it does not make it cease to be.

We can now see that the model of particularist moral reasoning developed in the last section was incomplete in one important respect—it omitted exclusionary reasons. The excellent ethical deliberator begins, as I said, by identifying the first-order reasons which are present in his choice situation, and their valences. But he must then determine whether any of the interests that ground those reasons also ground exclusionary reasons. And if they do, he must work out the width of the exclusionary scope of those reasons. We might model this process by beginning with the hypothesis that every interest grounds an exclusionary reason of scope 0, and then revising that hypothesis in the light of observations of features which determine the scope of an exclusionary reason in a particular situation—features like the ones enumerated and discussed in my account of the varieties of moral duty. These features would play a role analogous to that played by intensifiers with respect to first-order reasons. Once the deliberating agent has reached a conclusion about the scope of the exclusionary reasons present, he can turn his attention to the first-order reasons that remain unexcluded. At this point he will consider the weights of these remaining reasons, and update his judgments on the basis of observations of intensifiers and attenuators.

[20] I think it plausible that the existence of rules—both legal and customary—makes it psychologically easier for the agent to restrict his field of vision, as it were, with regard to his choice situation, and consider only those actions (and their outcomes) which fall within the bounds of the rules. This is one of the main sources of the social utility of rules.

There is a sound basis in Aristotle's own works for regarding recognition of and respect for one's own duties and the rights of others as a hallmark of the *phronimos*. As Fred D. Miller has convincingly shown, the concept of a claim-right was a familiar one to the ancient Greeks and was recognized by Aristotle. One of the clearest pieces of evidence for this comes from Aristotle's discussion of citizenship:

> [N]or are those persons [citizens] who partake of rights ['οι των δικαιων μετεχοντεσ] to the extent of undergoing and bringing lawsuits, for this also belongs to those who have a community as a result of treaties… (*Politics* III.1 1275a8–11, translated in Miller 1995, p. 98)

Here, we must translate *dikaion* as "right," or more precisely as "claim-right": claim-rights are what are pressed and disputed by those acting as plaintiff and defendant in a lawsuit—in ancient Greece just as in the modern West. In fact, as Miller shows, Aristotle recognizes all four species of Hohfeldian rights, and has a different term for each of them. Far from lacking a concept of individual rights, Aristotle avoids the modern confusion which results from having a single term that refers to all of these four distinct notions. Much of the dispute regarding the place of rights in Aristotle's moral and political theory results from the fact that he lacks such an umbrella term.[21] Nonetheless, Aristotle sees claims, liberties, powers and immunities as a closely related family of concepts (Miller 1995, pp. 107–108).

Phronesis is, fundamentally, excellence in reasoning about how to live a good life. Excellence in means- and ends-deliberation is an aspect of *phronesis*, since a good life is a life of achievement in a context of freedom. Excellence in ethical deliberation, at least to the extent of recognizing one's duties and the rights of others, is another aspect of *phronesis*, since a good life is a social and communal life. Miller has argued, on the basis of a set of passages from the *Politics*, the *Nicomachean Ethics*, and the *Eudemian Ethics*, that the virtue of *phronesis* directs us to live in a society in which we respect each other's rights, and in which legal rights are based on moral ones (Miller 1995, pp. 108–111, 128–139). The argument he reconstructs is based on the social nature of the good human life, and is similar to the Aristotelian argument I give above for the claim that those interests whose satisfaction is necessary for society to flourish ground perfect duties. To recognize the reasons that constitute those duties—or indeed the reasons which constitute the weaker duties—and act virtuously as a result of that recognition, is to exercise *phronesis* in its ethical aspect. My Neo-Aristotelian position thus goes farther than Aristotle's own view (at least as interpreted by Miller), and sees *phronesis* as directing us to recognize and respect our duties to others, even those not correlated with rights. I think it is fair to say, however, that Aristotle sees the sphere of the law, and so the sphere of exactable duties, as being significantly broader than I do. He would likely be comfortable with the idea that the State has a right that its citizens fulfill many of the duties I have identified as not justifiably exactable. So the real difference between my view and Aristotle's may be that mine is the more politically liberal—which should hardly come as a surprise.

[21] In Aristotle's writings, liberty = 'εξουσία, power = κύριοσ, and immunity = ἄκυροσ (Miller 1995, pp. 101–106).

Let us move on and consider the agent who is contemplating a choice within the limits of his duty. It may still be the case that his ethical deliberation and his deliberative preferences point him in different directions. This is the case of an agent who must choose between the morally permissible pursuit of his own ends, and the supererogatory pursuit of the interests of others. What makes the problem so acute is that we seem to have two modes of reasoning: on the one hand, the decision-theoretic mode of deliberation among actions whose potential outcomes are ordered in a preference-ranking determined by ends-deliberation; and on the other hand, the particularist mode of deliberation, which also aims to determine a choice among actions on the basis of the reasons that count in favor and against them in a specific case. And so we have reached the point at which these two modes of deliberation must be reconciled, must be made to work in concert with one another to determine a choice of action.

There are a few claims we will have to preserve in order to devise a satisfactory solution to this problem. Suppose the agent ranks e_1 (satisfying one of his own important interests) above e_2 (satisfying one of his friend's interests) in his deliberative preference-ranking, but now finds himself in a situation in which both are available, and the balance of reasons favors e_2. The first claim we want to preserve is that up until this point, the agent has deliberated over ends as well as anyone could. The second claim is that moral reasoning is practical in nature. It is not merely speculative reasoning about what one ought to do, which serves only to satisfy one's curiosity about the normative. It is reasoning about what to do. And so it must be possible for the agent, as a result of his moral reasoning, to decide to do what the balance of reasons favors. The third claim is that the agent's actions reflect his preferences over actions: the action he performs will be the action he prefers to preform, and that preference in turn reflects his judgments on how choiceworthy the outcomes are (as determined by ends-deliberation), and his judgments of how efficacious are the available actions in bringing about those outcomes (as determined by instrumental deliberation).

Now that we see how tight is the corner we have painted ourselves into, we should note one way in which we should *not* try to solve this problem. We should not simply say that the ethical deliberations of the practically wise agent will have the effect of leading him to reverse his preference between e_1 and e_2. For that way madness lies. Were we to make that claim, we would no longer be able to view preferences as even moderately stable. For consider the next choice situation the agent encounters in which he must once again decide between contributing to the satisfaction of that same important interest of his, or satisfying the same interest of his friend. In this next situation, however, the particular circumstances are such that the balance of reasons favors satisfying his own interest. And thus his supposedly deliberative preference-ranking is tossed to and fro on the waves of circumstance.

In order to solve the problem, then, we must see ethical deliberation as the predecessor to a new round of ends-deliberation.[22] If ethical deliberation is to be genu-

[22] This is a thoroughly Aristotelian claim. For Aristotle, a particular action is, in its own way, an ethical starting point (*arche*), just as much as the correct conception of *eudaimonia* is. The particu-

inely practical, it must influence the agent's choice of action *through* exerting an influence on his value judgments regarding his ends. But it must exert this influence in some way other than by prompting a simple reversal of preference within the agent's ranking of ends. Instead, and in keeping with the assumption that the agent entered the choice situation with preferences which were the result of excellent ends-deliberation up to that point, ethical deliberation prompts the agent to engage in a new round of *specificational reasoning*, as this process has been described in Chap. 4. The practically wise agent sees every choice situation as an opportunity to refine his system of ends, and thus to continue to work toward his own complete and fully specified conception of a good life. The conclusion of his ethical deliberation prompts the agent to ask himself whether the ends he sees himself as currently choosing between are not underspecified. In particular, he asks himself whether there is not some further specification of his ends which would bring his preferences in line with the conclusions of his ethical deliberation. Such a specification is always possible. The agent in our example will ask himself whether he is, and has been all along, committed to the achievement of e_1 *with* or *without* regard for the specific combination of considerations which, in this case, make pursuing e_2 the right thing to do. These are two ways of further specifying the end e_1. The agent will then engage in ends-deliberation regarding these two newly specified ends, and update his deliberative preference-ranking on the basis of that ends-deliberation. The reasons he considered during his ethical deliberation will now serve as the evidence which forms the basis for updating his probability judgments about his newly specified potential preference-rankings, and the learning experiences through which he uncovered the reasons which featured in his ethical deliberation may have led to a change in his taste for those reasons, which will in turn lead to an update of the desirabilities of his newly specified potential preference-rankings.[23]

The truly excellent ends-deliberator will, in a scenario like this, update his judgments and his desirabilities, and thus his deliberative preference-ranking, so that the conflict is eliminated: when he proceeds to deliberate instrumentally with his new preference-ranking over ends, the action to which he assigns the greatest expected value will now be the same as the action he judges he ought to perform.[24] This sort

---

lar action one has concluded is right in a concrete case is the datum that initiates a new round of practical inductive reasoning, which leads to a further refined and detailed conception of *eudaimonia*. Even the *phronimos* continues to learn about what it is to lead a good life whenever he encounters a new situation, deliberates about it, and acts (Reeve 2013, p. 234).

[23] First-order reasons for action thus serve as a basis for revising one's evaluations of the potential outcomes of one's actions, unlike social rules which, as we saw in the last chapter, leave those evaluations intact while conferring additional value on one course of action or another.

[24] The agent who falls short of reevaluating the outcomes of his potential actions in this way may be helped by the presence of a social rule, if that rule confers additional value on the course of action which he ought to perform. This is another source of the social utility of rules. If, on the other hand, a social rule confers value on a course of action which is not the one the agent has determined he ought to perform, the reasons which support breaking the rule serve as a basis for attempting to revise the strength of one's disposition to follow the rule in the present case (just as they serve as a basis for revising one's evaluations of one's potential outcomes). In the latter type of case, the excellent practical reasoned will successfully revise the strength of his rule-conform-

of continuous refinement of preferences, unlike the outright reversal we considered above, is consistent with the requirement that preferences be moderately stable. The agent still has a preference for satisfying his own important interest over his friend's interest; what he now recognizes is that he is not in a situation in which there is an action available which would *count* as satisfying that interest, given the way he has specified his commitment to it. If, the next time he must choose between these interests, the balance of reasons is on his side, no further change to his preferences will be prompted.

Here we see again why *phronesis* is excellence in ethical deliberation as well as excellence in means- and ends-deliberation. Continued excellent ends-deliberation—continued refinement of one's system of ends in response to the stream of evidence from newly encountered situations that bears on one's value judgments—depends on excellent ethical deliberation. The morally weak agent fails to alter his preferences, intentions, and actions in light of this evidence. His ethical deliberations fail to prompt a change of mind, and he persists in preferring an action other than the one he ought to perform. This form of moral weakness is the flipside of weakness of will, to which I alluded in Chap. 7.[25] It is a failure to change one's intention when one should. But note that this sort of weakness is consistent with our commitment to a mild form of both judgment- and reasons-internalism. For it need only be the case that the agent's appreciation of his particular reasons for action has some effect on the value judgments and desirabilities which determine his preference-ranking over ends—not that it have a decisive effect—for mild judgment-internalism to be preserved. And likewise, it need only be the case that the agent could have developed into someone for whom that effect would have been decisive, for reasons-internalism to be preserved.

4.4 Eudaimonia *and* Iustitia

One way to interpret my reconciliation of deliberative preferences and the recognition of one's moral duty is as a eudaimonistic appropriation of Duns Scotus' distinction between *eudaimonia* and *iustitia*, or well-being and morality, which Scotus

ing disposition in that case so that the action he values most highly is the one he ought to perform.

[25] Since ethical deliberation is the closest thing in my theory of practical reasoning to Aristotle's own conception of deliberation, this phenomenon of moral weakness is even closer to his conception of weakness of will in *NE* VII than any of the forms of weakness of will I discussed in Chap. 7. One symptom of moral weakness is thus the tendency to engage in post-hoc rationalization for one's choices. This phenomenon can be modeled as the agent, who has failed to revise his preferences through specificational reasoning on the basis of ethical deliberation, returning to that very ethical deliberation and "cooking the books"—revising his judgments about the weights of his reasons so that his conclusion about what he ought to do aligns with his preference about what to do. The agent may of course not be aware that he is doing this.

himself took as undermining any eudaimonistic approach to ethics.[26] Scotus saw *eudaimonia* as the object of one of the inclinations of the will: the *affectio commodi,* or inclination of advantage. For Scotus, this was the will's *natural* inclination, just as (in Aristotelian physics) the natural inclination of a stone is to fall toward the center of the Earth. For this reason, Scotus denied that an account of *eudaimonia* could serve as the basis for morality (*iustitia*). Morality is only possible where there is the possibility of free action. And for Scotus, to be able to act freely just is to be able to act without regard for one's natural inclinations. So the very possibility of morality rests on there being a second inclination of the will—on the will being able to act without regard for its natural inclination. This second inclination cannot be an inclination to act in a way that is arbitrary, since that is the wrong kind of freedom required by morality. Moral action, for Scotus, is action in accordance with divine law, the content of which is known through revelation. So for moral action to be possible, the will must have an inclination to act without regard for the agent's well-being—and thus be able to act freely—but at the same time to act in accordance with divine law. The second inclination of the will, which makes both freedom and moral action possible, is thus the *affectio iustitiae*, the inclination of justice.

One way to interpret the neo-eudaimonistic project is as expanding the notions of *eudaimonia* and *phronesis* so that they encompass morality as Scotus understands it. Corresponding to the cognition of the good that prompts the will to act in accordance with the *affectio commodi*, my theory has the exercise of *phronesis* in means- and ends-deliberation. Corresponding to the cognition of the revealed law of God that prompts the will to act in accordance with the *affectio iustitiae*, my theory has the exercise of *phronesis* in ethical deliberation. On the one hand, the neo-eudaimonist takes reasons, the object of ethical deliberation, to be grounded on interests—aspects of well-being. On the other hand, the good life, *eudaimonia*, is not simply defined as the life of achieving valuable ends. It is also a moral life, a life of achievement within the bounds of moral duty. What I reject is Scotus' claim that freedom and morality are only possible for the agent capable of acting without regard for his well-being. Instead, I reconcile ethical deliberation and ends-deliberation, and claim that for the practically wise agent, his recognition of what he ought to do guides a process of refinement of his judgments of the value of his ends, a process which ultimately contributes to determining his choice of action. It is just that sort of responsiveness to reasons which makes moral action possible.

Scotus' position is likely to strike the reader as surprisingly similar to Kant's, and indeed, I think it is fruitful to interpret Kant's ethical theory as an attempt to both secularize and radicalize Scotus' theory. Kant radicalizes Scotus' notion of freedom, by understanding free action not merely as action not motivated by desire (Scotus' natural inclination), but as action not performed in accordance with any law which the agent does not determine and legislate for himself. Free action must be autonomous, in the sense of "self-legislated" which is Kant's interpretation of the term, rather than heteronomous. This way of radicalizing Scotus' notion of freedom obviously implies a secular departure from his theory.

[26] See in particular his *Ordinatio* Book 2 Distinctions 6 and 39, and Book 3 Distinctions 17 and 26.

The practically wise agent of neo-eudaimonistic theory is autonomous in the Kantian sense, as well as in all the other senses discussed so far. Just as he satisfies the competence and authenticity aspects of autonomy through his excellence in means- and ends-deliberation, and the self-control aspect, he satisfies the self-legislating aspect through his excellence in ethical deliberation. He uses his practical reason to determine for himself the presence, valence, and weight of the reasons in each situation in which he acts. And so we may add self-legislation as a third aspect of the rational dimension of autonomy, and understand it in terms of excellence in ethical deliberation, which we have seen is an aspect of *phronesis*. We can accommodate the Kantian notion of autonomy within the context of the basic neo-eudaimonistic claim that the *phronimos* is the autonomous agent *par excellance*. However, insofar as Neo-eudaimonism sees the autonomous moral agent as engaged in a communal project, stretching across history, to refine and calibrate our individual conceptions of the normative web—though without acting on any such conception which he has not been rationally persuaded to accept—the theory moves beyond the Kantian notion of self-legislation in a Hegelian direction.

Kant also radicalizes Scotus' notion of morality, and he does so in a way which may seem troublesome for both the Scotian and the neo-eudaimonist. For Kant, morality is radically categorical. This idea embraces two claims. The first is that moral requirements apply to the moral agent regardless of his ends. This is a form of reasons internalism, and both the Scotian and the neo-eudaimonist embrace it—though the latter only within the bounds of the Principle of Minimal Psychological Realism. The fact that Scotus sees morality as grounded by God's will, and the neo-eudaimonist sees moral reasons as grounded by human interests, does not make these theories any less categorical in this sense. One need not actually care about God's will, or about the interests of others (or even of oneself), in order for moral requirements to apply to one. They apply categorically. For Kant, however, morality also commands categorically. The rational will cannot comprehend its moral duty without either acting in accordance with that duty, or ceasing to be a rational will altogether. This is a very strong sort of judgment-internalism. As I mentioned above, it is of course psychologically possible for a moral agent to disregard what he recognizes to be the exclusionary reasons that apply to him. To act on these reasons, to confine one's decisions to those actions that are within the bounds of one's duty, is a voluntary choice. Likewise, it is possible for one to disregard revealed divine law. Even one who has faith in that law may fail to act on that faith. Kant is searching for an understanding of morality that makes this sort of comprehension-*cum*-failure impossible.

The neo-eudaimonist, as we have seen, rejects this sort of strong judgment-internalism. He sees value judgments and judgments about reasons as always exerting an influence on an agent's reasoning about what to do, but does not take that influence to be necessarily decisive. But a remnant of the Kantian challenge remains to be dealt with. The neo-eudaimonist must provide an account, where the Kantian need not, of how and why we sometimes succeed in acting according to these judgments, and sometimes fail. But we already have such an account in place. Our ability to act on these judgments is, as already discussed, liable to be diminished by a

variety of disabling and interfering environmental and social cues, as well as by physical weakness and depletion. Success in acting on our reasons and values requires exercising a well-developed capacity for self-control; and we fail when we encounter circumstances adverse enough to overwhelm this capacity.

4.5 Moral Institutionalism

The most powerful challenge to moral particularism that I know of comes from Joseph Heath's institutional theory of morality (though Heath does not explicitly use his theory to argue against particularism). According to the institutional theory, "Social norms are the 'ontological' correlate of moral judgments. Morality is 'about' the rules that govern our interactions" (Heath 2008, p. 285).[27] If moral judgments are about rules, it is certainly hard to see how one could engage in healthy moral reasoning without appeal to rules. Heath's argument for this position counts, I believe, as one of the most sophisticated and ingenious projects in meta-ethics and moral psychology of the last quarter century. But I also believe there is sufficient reason to doubt that it succeeds.

I begin with a brief presentation of the highlights of Heath's argument. The first crucial move is a particular interpretation of the "linguistic turn" in twentieth century philosophy:

> Rather than treating the intentionality of consciousness as primitive, philosophers began to consider the possibility that semantic intentionality might be more fundamental (or perhaps equiprimordial). After all, insofar as our thoughts have content, they can also be given linguistic expression. Thus the set of intentional states is also a set of states with propositional content. The suggestion at the heart of the linguistic perspective is that the intentionality of these mental states may be inherited from the propositions that give them their content…It is not difficult to imagine the mechanism that might be responsible for such an order of explanation. People who talk to themselves as they try to resolve a problem are often described as "thinking out loud." It is possible that the opposite is true—that thinking (in the sense of rational, analytic thought) is really a form of silent talking. (Heath 2008, p. 101)

According to Heath, the challenge for the philosopher who has taken the linguistic turn then becomes that of finding a way to account for the development of language without assuming the prior development of intentional mental states. The first step, since language is a rule-governed activity, is to account for the origin of normativity:

> In order to understand language we need an account of normativity grounded in a theory of behavior. In other words, we need to explain, in a way that does not presuppose intentionality, how it might come to be the case that certain actions are right and others wrong (or correct and incorrect). Naturally, such an account will not initially account for how agents are able to say, or to believe, that certain actions are either right or wrong. It will only

[27] Heath is not the only philosopher to argue for an institutional theory of morality. See also (Binmore 2011; Gauss 2011). I focus on Heath because I think his argument is the best, largely owing to his explicitly tackling the problem of the originary emergence of normativity.

> explain what it is for agents to treat certain actions as right or wrong, in their conduct. It will be an account of what Robert Brandom calls "norms implicit in practice." (Heath 2008, p. 111)

The paradigmatic example of a behavior which is norm-implicit is sanctioning behavior. Heath conceives of the process by which normativity arises "out of the primordial non-discursive ooze" (Brandom 1994, p. 626) as that of *reciprocal sanctioning* between two agents, the first of whom acts, while the second observes the action:

> The second agent has what might be called an "expectation of behavior"—she expects the first to behave in a certain way. If the first person anticipates these expectations, he may develop what we can call an "expectation of recognition"—he expects her to respond correctly to his actions, to punish him only when it is appropriate to do so, or to reward him when he is entitled to it. Whenever either expectation is disappointed, sanctions are imposed. In this way, the second person's sanctioning efforts become subject to sanctions by the first, just as the actions of the first are subject to sanctions by the second. (Heath 2008, p. 114)

Reciprocal sanctioning involves a virtuous circle—the second agent expects that the first has an expectation of recognition, the first expects the second agent to have this expectation, and so on—and it is from this practice that normativity originally arises:

> [W]hen sanctioning is reciprocal, two agents can each act in a way that confers normativity on the actions of the other, and this, by extension, confers normativity back on their own actions. There is nothing left in the interaction that could count as "mere behavior." In particular, because everyone engages in a normative assessment of everyone's conduct, everyone has no choice but to adopt such an assessment of his or her own conduct (at least implicitly). Thus it is plausible to suggest that "original normativity" inheres in the practices of a community in which everyone sanctions everyone else, and sanctioning conduct is itself sanctionable conduct. Furthermore, there is no reason to think that this account presupposes cognitive abilities that are beyond the reach of prelinguistic hominids. The sanctioning behavior can be described as a set of responsive dispositions—it need not at any point involve any contentful representation of what the other has done, or will do. (Heath 2008, p. 115)

Of course, we already have good reason to resist the characterization of the assessments and actions of pre-linguistic hominids as "normative." The behavior in question can certainly not be modeled as purely instrumental. It must be modeled on the assumption that the individuals make non-instrumental evaluative assessments of the kind discussed in the previous chapter, according to which there is value in conforming to a rule beyond the value of the outcome of the action which counts as following it (so that rules will be followed even when doing so is costly to the follower from a purely self-regarding standpoint), and (reciprocally) value in the action of punishing a rule-breaker beyond the value of the outcome of that action (so that rule-breakers will be punished even when doing so is costly to the punisher from a purely self-regarding standpoint). But there is no need or justification for an interpretation according to which these individuals are, or are acting as if they are, taking the rules implicit in their practice to be reasons, and engaging in normative assessment and behavior as we have understood these notions. But we can bracket this

point for the moment and continue with the account, since the emergence of the sort of rule-conformative behavior Heath describes will prove to be an important step on the way to the emergence of normativity.

The next claim is that the emergence of normativity (in Heath's sense of strongly reciprocal general rule-conformism) makes possible the development of culture, which serves as the basis for the development of language. From the development of language flows the emergence of intentional states, human intelligence, and altruism:

> Once culture dependence is established, in the form of a norm-conformative disposition (imitative conformity coupled with moralistic punishment), one can then explain the emergence of propositionally differentiated speech (as Brandom's pragmatic theory of meaning shows), which can in turn be used to explain the origins of mental content, intentional states, and finally the intentional planning system that is at the root of our superior practical intelligence. Finally, it is much easier to see how altruism (and ultimately, cooperation) could persist as a culturally transmitted pattern of behavior…Thus norm-conformity appears to be the key that opens all the locks. We are not just intelligent creatures who happen to like following rules; rather, following rules is what makes us the intelligent creatures that we are. (Heath 2008, p. 201)

The upshot of this argument is that human practical rationality is primarily a matter of reflecting on and applying rules, on the one hand, and criticizing rules by appeal to a rich background of further rules, on the other. Thus, a theory of the origins of normativity, culture, language, and intentionality gives rise to the understanding of moral judgment expressed by institutionalism.

The source of the difficulty with Heath's argument is that it is shot-through with a confusion he inherits from Brandom. Brandom does not have one pragmatic theory of meaning. He has two theories, of very different types. The first, which usually goes by the name of inferential-role semantics, is a semantic theory of meaning. It is a theory which aims to provide the general features of any adequate answer to the question "What is the meaning of this expression?" Brandom is a pragmatist insofar as, following Sellars and Wittgenstein, he does not conceive of meaning as a *relation*. He is not, in other words, developing a theory which aims to identify some class of entities—whether abstract objects, mental representations, or what have you—which can be matched with words and serve as their meanings. Wittgenstein zeroes in on this idea as one of the pervasive mistakes in twentieth century philosophy with his characteristic humor:

> You say: the point isn't the word, but its meaning, and you think of the meaning as a thing of the same kind as the word, though also different from the word. Here the word, there the meaning. The money, and the cow that you can buy with it. (Wittgenstein 1953, p. §120)

A pragmatist version of a semantic theory of meaning does not seek to identify any class of entities as the meanings of words. Rather, it understands the notion of knowing the meaning of a word or expression in terms of knowing the uses to which the expression is put within a language or, more generally, a form of life. Inferentialism is a particular type of pragmatist theory of meaning, which focuses on the inferential roles played by the expressions in a language.

But Brandom also advocates a theory of meaning of a very different type: an ontogenetic and phylogenetic theory. This is a theory about the *development* of language and linguistic capabilities. And it is a theory according to which the development of language *precedes* the development of intentional mental states. The fact that one is a pragmatist with respect to one's semantic theory, however, does not commit one to an ontogentic or phylogentic theory in which language development precedes the development of intentional states.

Sellars, in whose footsteps both Brandom and Heath take themselves to be following, certainly subscribed to a use-based or function-based analysis of the semantic properties of language. As far as his semantic theory is concerned, he is a pragmatist. And this commitment is key to his argument against the Cartesian version of the Myth of the Given, according to which our psychological states are given to us and play the role of unjustified justifiers, on which our knowledge of the external world can be founded:

> The argument presumes that the metalinguistic vocabulary in which we talk about linguistic episodes can be analysed in terms which do not presuppose the framework of mental acts; in particular that
>
> "…" means p
> is not to be analysed as
> "…" expresses t and t is about p
> where t is a thought. (Sellars 1957, p. 522)

We must, however, remember that (as Willem deVries puts it) "[Sellars'] *philosophical* interest in mind is a metalevel interest in understanding our *concepts* of mind, not an object-level interest in particular minds or even generalizations about minds" (deVries 2005, p. 172). In order to undermine the Cartesian version of the Myth of the Given, Sellars sets out to show that it is possible for there to be a community of language-users who lack any psychological concepts, but who go on to develop such concepts. His argument, the centerpiece of which is the "myth of Jones," is that psychological concepts can be developed by positing a set of internal states with properties that are analogous to the properties of speech acts.[28] The crucial point here is that insofar as Sellars subscribes to any developmental theory at all, it is an *epistemic* one: a theory about how our knowledge and understanding of our intentional states develops. It is not an ontogenetic or phylogentic theory. There is no implication that intentional mental states emerge as a result of the development or acquisition of language. Rather, Sellars' point is that language-users' understanding of their intentional mental states is modeled on their understanding of their linguistic behavior. In fact, in his mature thought, Sellars explicitly endorsed the hypothesis that the development of intentional, representational states precedes the development of language:

> In the domain of the mental, language is prior in the order of knowing. What, then, in this domain, is prior in the order of being? There is no easy answer to this question, for I can think of no simple way of putting it which is not misleading. Perhaps the best way of putting

[28] The myth of Jones is introduced in (Sellars 1956); for Sellars' mature presentation of the process of constructing a theory of intentional states, see (Sellars 1968, §23).

> it, of indicating the general character of the answer, is by saying bluntly, *animal representational systems*…My disagreement with the classical view [of the relationship between language and thought] takes its point of departure from the fact that I construe *concepts* pertaining to the intentionality of thoughts as derivative from *concepts* pertaining to meaningful speech. (Sellars 1981, pp. 326–327)

Sellars goes on to claim that in pre-linguistic representational systems, representational states have *propositional form*; they represent objects, and represent them *as of* a certain character (Sellars 1981, p. 336). Language users, however, are, in virtue of their acquisition of language, in possession of a representational system whose representational states have *logical form*. Such systems can explicitly formulate and obey rules of inference. The operations of representational systems that do not do this, however, are still appropriately described by *mentioning* logical operations (Sellars 1981, p. 340). Their operations produce inferential patterns, which are "uniformities in the occurrence of representational states" (Sellars 1981, p. 337). The development of representational systems with propositional but not logical form thus serves as a bridge to the development of systems with logical form. As deVries puts it, the former "'ape reason' by possessing dispositions to change their representational states in ways that parallel the inferential changes made by a fully-fledged reasoned acting on general [logical/inferential] principles" (deVries 2005, p. 189).

The point of this brief excursus into Sellars' thought is to demonstrate, from a philosophical and intellectual-historical point of view, that there is nothing inherent in the linguistic turn, or even in the post-turn pragmatist tradition, which commits one to an ontogenetic or phylogenetic theory according to which the development of normativity, culture, and language precede the development of intentional representational states. There is nothing inherent in modern pragmatism which would lead us to believe that, as Heath put it, "we need to explain, in a way that does not presuppose intentionality, how it might come to be the case that certain actions are right and others wrong." Nor is there reason to hold that the "cognitive abilities of prelinguistic hominids" do not extend to "contentful representation of what the other has done, or will do." The ontogenetic and phylogenetic theses of Brandom and Heath are further claims, whose denial is perfectly consistent with a commitment to Sellarsian pragmatism.

The mere fact that these claims about the development of language and intentionality are not implied by an analytic pragmatism like Sellars' is not an argument against them—and nor is the fact that Sellars himself held contrary views. As ontogenetic and phylogenetic hypotheses, however, the place to look for evidence for or against them is in the developmental sciences: developmental psychology and neuroscience, and evolutionary biology and anthropology. But the latest developments in these fields do not support the hypotheses of Brandom and Heath. Tyler Burge has recently argued, at great length and in admirable detail, that the results of contemporary psychological and neuroscientific research speak unequivocally against the idea that any sort of linguistic capacity is required for an organism to be capable of producing objective representations of its environment (Burge 2010). It is doubtful, moreover, that the sorts of *expectations* Heath attributes to pre-linguistic ani-

mals—the expectation of behavior and the expectation of recognition—are possible without intentional states. The view in contemporary cognitive science is that the precursor to the ability to have expectations of these sorts is the ability to represent goals (Castelfranchi and Lorini 2003). There is extensive evidence that chimpanzees, who engage in cooperative and sanctioning behavior, not only have goal-representing intentional states, but understand that other beings have them as well and use them to guide their behavior (Call and Tomasello 2008). That is to say, chimps are capable not only of representing the world—of having representational states—but also of a rudimentary sort of representational *thought*: of interpreting their representations as representations, and attributing both representational states and the ability to interpret representations as representations to others. This is what enables them to cooperate, and to engage in genuine sanctioning behavior, as opposed to merely responding aggressively to behavior they observe and do not like. Cooperation takes place for the sake of achieving goals. The chimp who cooperates recognizes that the one he cooperates with is in pursuit of one of his goals, and he cooperates for the sake of furthering the pursuit of one of his own goals. Non-cooperative behaviors—actions by one chimp which interfere with another chimp's pursuit of his goals when co-operation would benefit the first chimp, free-riding behavior, or defection from co-operation—are sanctioned by the one who has experienced the interference or defection, or whose efforts the free-rider has taken advantage of, in return for being frustrated in the pursuit of his goal. The strongly reciprocal cooperation which is the precursor to early systems of rules—shared standards governing sanctioning—thus assumes the development of the sort of representational capacities that can be attributed to pre-linguistic animals. These same representational capacities are likewise the precursor to early systems of values—shared standards governing the acceptability of potential goals. At the primordial origin of both normativity and value we find the capacity to represent commitment to the achievement of goals—hardly a surprise for the neo-eudaimonist.

Most significantly for my purposes, there is an important gap in the Brandom/Heath account of the transition from the appearance of strongly reciprocal cooperation (i.e. cooperating with those who are cooperative in turn, and punishing those who are not, even when doing so is costly from a purely self-regarding standpoint) to the development of human culture and the emergence of normativity (and thence, through the development of language, to long-term planning, human intelligence and rationality, and human hyper-sociality). The dominant—though by no means uncontroversial—hypothesis in evolutionary anthropology, owing largely to the work of Michael Tomasello and his colleagues at the Max Planck Institute, is that all of these developments were made possible through the development of a capacity for *shared intentionality*: the ability and the disposition to share experiences and goals with others, and to represent to oneself the differing perspective of another on a shared situation, including the place of oneself in it (Tomasello 2001, 2009, 2010, 2014; Tomasello et al. 2005). As Josep Call explains, "[S]hared intentionality is responsible for the appearance of a suite of behaviors, including joint attention, declarative communication, imitative learning, and teaching, that are the basis of cultural learning and the social norms and traditions present in every human cul-

ture" (Call 2009, p. 368). It is the capacity for shared-intentionality, rather than rule-conformism, that is the "key that turns all the locks."

An appreciation of the foundational role of shared intentionality opens up an alternative interpretation of the roles of culture and rules in human development. Rather than being primarily a mechanism for the transmission of social norms, culture is a shared, inheritable, value-based system for structuring individual and communal interests and goals, and transmitting understanding of and commitment to those goals. The (unintentional) transmission of rules-implicit-in-practice is an important part of this process; but explicit rules and their usefulness are a later cultural discovery. Greater social and cultural achievements are possible for creatures who have internalized explicit rules. Of course, this cultural discovery can only be exploited by creatures who have already evolved a general rule-conformative disposition. But just as we have found both philosophical and empirical reasons to believe that the evolution of language follows, in the order of being, the evolution of mental representation, we will now see that there is empirical reason to believe that the evolution of a general rule-conformative disposition follows the evolution of a capacity for shared intentionality, which in turn presupposes both a disposition for strongly reciprocal cooperation and a capacity representational thought.

Both Heath and Gaus emphasize the fact that human beings are disposed not merely to engage in reciprocal cooperation and punishment, but to follow all sorts of rules and punish all sorts of rule-breakers (Heath 2008, p. 184; Gaus 2011, p. 112). This is perfectly correct. But this fact does not show that rule-conformism is as old, evolutionarily, as strongly reciprocal altruism, or that the latter is an instance of the former. The evidence belies these claims. The display of strong reciprocity across multiple contexts—including those, such as the ultimatum game, which cannot be described as instances of cooperation—is precisely what we do *not* observe among our closest evolutionary relatives. Chimpanzees are strongly reciprocal cooperators: they are disposed to cooperate with those who reciprocate cooperation, and to punish those who defect, even when doing so is costly (Boesch 1994; Muller and Mitani 2005). But they behave like perfect individual utility maximizers in non-cooperative contexts such as the ultimatum game (Jensen et al. 2007, pp. 107–109). What chimps lack is shared intentionality. Unlike humans as young as 1 year old, they do not exhibit the behavioral signs of sharing goals. As a result, even their cooperative behavior is individualistic: as already described, one chimp cooperates with another in order to attain *his own* goal. Chimps do *not* cooperate for the sake of attaining *shared* goals (Tomasello and Carpenter 2007, pp. 121–125; Call 2009, pp. 372–372). Shared intentionality helps bridge the gap from reciprocal cooperation to robust reciprocal rule-following.

To better understand how shared intentionality helps to bridge this gap, consider the practice which marks the emergence of normativity and the first instance of general, implicit rule-conformative behavior: the practice of reciprocal sanctioning This practice is inaccessible for creatures that lack shared intentionality. This is so even for strongly reciprocal cooperators, who possess representational states and attribute such states to others (as they must if they are to have the necessary expectations of behavior). Without shared intentionality, chimps are incapable of recogniz-

ing that another creature has a false belief about the world, and of predicting what that other will do based on that false belief—abilities possessed by human infants (Call and Tomasello 2008, pp. 190–191; Tomasello and Moll 2013). A chimp who has cooperated is *incapable* of understanding another chimp's aggression as a sanction resulting from a false belief that he has not done so, and of distinguishing that action from any other form of aggression. The first chimp does not have an expectation of recognition, if this means an expectation that cooperative behavior will not be followed by sanctioning behavior. He cannot conceive of aggression that follows cooperation as a sanction. He simply has an expectation that cooperative behavior will not be followed by aggression. And so when the first chimp responds to the second with further aggression, this is a response to what seems to him to be unprovoked aggression. We can describe that response as itself a sanction, but not as a counter-sanction—not as a sanction of a sanctioning act. It is a sanction of an act of unprovoked aggression. The very possibility of counter-sanctioning, and thus of a community in which sanctioning behavior is itself sanctionable, depends on the ability to recognize an act as one of unjustified sanction. This is what the chimps lack. To describe their behavior as reciprocal sanctioning is to interpret it from our own perspective, while forgetting that what makes that interpretation possible is the very shared intentionality the chimps lack. Their potential for implicit normative assessment, for genuine sanctioning behavior, is exhausted at the end of the first round. Reciprocal normative assessment is out of their reach. The genuinely reciprocal sanctioning of pre-linguistic hominids is made possible by the evolution of a capacity for shared intentionality; normativity cannot emerge prior to this.

The observation that rule-recognition and rule-responsiveness is phylogenetically posterior to shared intentionality is recapitulated ontogenetically. Having developed the capacity for goal-directed behavior in their first year, human infants between 12 and 18 months of age not only develop the capacity for shared intentionality, and the motivational disposition to share goals and experiences; they also develop the capacity *both* to evaluate actions as means to ends *and* to evaluate ends in themselves as good or bad (Rakoczy et al. 2008, p. 875). The capacity to recognize and respond to social rules, on the other hand, does not begin to emerge until after 18 months, and is not significantly developed until 3 years of age (Rakoczy et al. 2008, pp. 879–880).[29] This is the point at which children begin to sanction others for harms done to third-parties which do not affect them, a behavior not found in non-human primates (Tomasello 2014, p. 87). The foundation of our abil-

[29] Marco Iacoboni has argued that what he calls the "normative view of intentionality," according to which our ability to have intentions and recognize the intentions of others is "determined by social practices," is supported by research that shows that when adults interpret the intentions of others in acting, it is the "social brain" that is engaged—in particular, the orbitofrontal cortex and the limbic system (Iacoboni 2003, pp. 130–132). But this research shows nothing of the sort, at least, not if what Iacoboni means is that the emergence of social conventions and rules is developmentally prior to intentions and intentional states. Indeed, the research of Rakoczy *et al.* cited above shows that this *cannot* be the case. Rather, what Iacoboni's research supports is the entirely uncontroversial claim that adults normally appeal to their knowledge of social conventions and rules in the process of figuring out what someone else's intentions are in acting.

ity not only to represent our own goals and the goals of others, but also to reason about means and to recognize the place of ends in a system of shared values, is laid prior to that of our ability to recognize, respond to, and reason with rules.

The evolved traits which are necessary precursors to the emergence of genuine normativity, in addition to the disposition for strongly reciprocal altruism and the capacity for shared intentionality, are the behavioral dispositions which are the precursors of the core virtues and the moral emotions such as shame and guilt.[30] These set the stage for proto-normative rule-conforming behavior. This is behavior that cannot be modeled as an aping of purely instrumental reasoning, and must be modeled as an aping of the sort of non-instrumental, rule-based reasoning and behavior discussed in the last chapter (though we should probably see the development of this behavior as a gradual process, beginning with the very rudimentary pre-normative behavior of the earliest strongly reciprocal cooperators).[31] This is the stage at which there are rules implicit in practice. It is also the stage at which shared systems of values emerge—first jointly held between pairs, and then collectively held within larger groups (Tomasello 2014, pp. 83–84). It is the beginning of human culture. Here lie the pre-linguistic hominids, who possess a general strongly reciprocal implicit-rule-conformative disposition, rather than the specific disposition for strongly reciprocal cooperation found in lower primates. Their practices are a bridge between lower primates and modern humans. For creatures who have evolved these traits, the evolution of linguistic ability ushers in full-blooded normativity by placing us in the logical space of reasons, giving us the ability to recognize and respond to reasons (including the fact of there being a rule) as reasons, and to see our behavior as reasons-responsive (Tomasello 2014, ch. 4).[32] It opens the door to explicit rule-conforming behavior—the ability to formulate the rules one follows. This is

[30] For the importance of virtue dispositions, see Gintis (2010, pp. 73–75); for the moral emotions, see (Bowles and Gintis 2011, ch. 11).

[31] To describe such behavior as "aping" the observance of rules is not merely an extension of Sellars' turn of phrase, but is in all likelihood how Sellars would describe it. For Sellars, there are no genuine rules until there are agents who are capable of recognizing and responding to rules as rules (Sellars 1974, pp. 423–424).

[32] The evolutionary anthropologist Terrence Deacon has made a fascinating argument that it was specifically the necessity of reciprocal altruism in mating—i.e. the self-restriction of sexual access to a single mate, in return for others doing the same—for the stability of early hominid communities, that provided the evolutionary pressure for early hominid brains to develop the capacity for symbolic representation, which he identifies as the foundation of human linguistic capacity. This initiated a virtuous spiral of brain-language co-evolution. See (Deacon 1997, ch. 12). It is worth noting that Deacon's overall account of the origins of language, as he presents it, implies that chimpanzees do not have a theory of mind, since he takes it that the capacity to represent another as minded requires the capacity for symbolic representation that he attributes uniquely to humans (Deacon 1997, ch. 13). It thus appears to be in tension with Tomasello's research. However, it is possible to revise Deacon's account in such a way that its major insights are preserved, while making it compatible with and complementary to Tomasello's. And there are good independent reasons for doing so (Stjernfelt 2012). In essence, Deacon may have actually identified the specific sociocultural situation that gave rise to the evolutionary pressure for hominids to develop shared intentionality, rather than the more basic capacity to represent another as minded. The former is the immediate pre-requisite for the sort of symbolic representation which is unique to humans, and

also the evolutionary stage at which sanctioning for harms done to third-parties likely emerged (Tomasello 2014, p. 87). The evolution of language thus opens a path to viewing one's own behavior, the behavior of others, and shared situations from a fully normative third-person, agent-neutral perspective, in addition to the second-person perspective of another individual with whom one is interacting directly, which is made available to pre-linguistic hominids by the initial emergence of shared intentionality. This marks the final step in a transition from a simpler form of shared intentionality—joint intentionality—to a more complex form—collective intentionality (Tomasello 2014, pp. 68–79, 113–123). The subsequent, post-linguistic development among humans of more and more sophisticated rule-conformative dispositions and cooperative behaviors, alongside more and more complex systems of social rules, is a paradigmatic example of gene-culture coevolution (Bowles and Gintis 2011, ch. 7; Gärdenfors et al. 2012). The full sequence of evolutionary development from mental representation to full-blooded normativity—from goal-representation and goal-directed behavior, to reciprocal cooperation, shared intentionality and shared values, conformity to rules-implicit-in-practice, language, and finally explicit-rule-conformism and third-party rule-enforcement—is, as noted above, recapitulated in the development of modern human children. As human culture continued to develop, norm-internalizing became the method of choice for transmitting cultural knowledge and values across generations (Scott 1971; Bowles and Gintis 2011, ch. 10).

From this perspective, social rules are tools; they develop, persist, change, and go extinct according to how well or poorly they support the cultural value system of the time. Those living within a culture, who have internalized the rules, do not see them as such, since to internalize rules is precisely to cease to see them merely as instruments deployed toward some further valuable end. But from the external perspective of the student of the phenomenon of culture, this is precisely what they are revealed to be. And sound cultural criticism, wherein we evaluate and revise the system of social rules in which we live, requires that we be capable of bracketing those rules we ourselves have internalized and scrutinizing them as instruments of social achievement.

The origins of normativity and value, of culture and language, and thus of human intelligence, sociality and civilization, are found in our capacity and disposition to share goals, to form "we-intentions" which can only be fulfilled if we fulfill them together. And so here too is where we should look to find a naturalistic basis for morality. Moral judgments, at the end of the day, are not about social rules. They are about our interests and goals, individually and collectively. They are about the value we attribute to these, and the reasons we take ourselves to have for pursuing them, both in general and in one particular way rather than another. When we judge social rules from a moral point of view—as we certainly do—we judge them not against a background of further rules, but against a background of reasons- and values-based judgments about the interests and goals which the social norms in question do or do

which the capacity for linguistic communication rests on, while the cognitive capacities of lower primates are sufficient for the latter.

not promote. Digging deeper into our moral background means questioning and judging these presuppositions, with an eye to moving ever-closer to a final conception of *eudaimonia* and the Normative Order. The story that emerges is one that supports Neo-eudaimonism.

Some normative judgments are essentially rule-invoking—grammatical judgments, for example. This is because language is essentially a rule-governed activity. Heath has persuasively argued that linguistic communication could only evolve among creatures with a general rule-conformative disposition—creatures whose behavior must be modeled in terms of treating rules implicit in practice as a source of value conferred on actions beyond the value attributed to their outcomes (Heath 2001, ch. 2–4). Given this, once we have evolved the ability to communicate linguistically, and with it the ability to recognize something as a reason, it is natural that the first things we recognize and treat as reasons for action are those very rules implicit in our practices (or our new explicit statements of them). But the fact that it is the development of language that provides our entry into the logical space of reasons, and that a general rule-conformative disposition is required for this development, does not entail that we must continue to identify reasons with rules or principles in all domains of inquiry even if they are the obvious candidates at first—it does not entail that we are forever entrapped in a 'prison-house of rules,' with no way to justify introducing, critiquing, revising or rejecting one rule except by appeal to another. The particularist contends that at least one body of normative judgments—ethical judgments—can perfectly well be made, argued about, and justified without appealing to rules. If a disposition to follow rules, even when merely implicit, really were the ground of normativity and rationality, there would be a strong presumption in favor of a universal rule-based theory of reasons. But we have seen that the empirical backing for such a presumption is lacking—the ground of full-blooded rationality and normativity, as well as for a general rule-conformative disposition and the capacity for linguistic communication, is the capacity for shared intentionality. So there is nothing inconsistent in simultaneously committing oneself to a naturalistic account of the emergence of normativity, and a particularist account of moral judgment and reasoning as being fundamentally about our individual and collective goals, the value we attribute to them and the reasons we have for pursuing them.

5 Conclusion

I have now completed my account of *phronesis*, excellence in reasoning about how to lead a good life. We can now see that it is excellence in all forms of practical deliberation: excellence in reasoning about means, ends, oughts, and musts (i.e. duties). I have explicated the relationship between ends-deliberation and ethical deliberation, and described the process of properly resolving conflicts between them. The *phronimos* is the one who possesses and exercises the virtue of *phronesis* in all its aspects—he is both the author of his own life and a self-legislator; one who

achieves authenticity by training his emotional attachments to track his value judgments; and one who, through his exercise of a well-developed power for self-control, acts on the conclusions of his practical reasoning even in the face of adverse circumstances and interfering environmental cues. He is thus the paradigmatically autonomous agent.

I have also completed my account of the good life—the life of flourishing, or *eudaimonia*. I had already characterized such a life as one of achievement of ends chosen autonomously in a context of freedom. I have now added the requirement that the pursuit and achievement of those ends occur within the confines of one's moral duty, and that one's life be characterized not only by achievement but also by the regular performance of virtuous actions. Virtuous action is something more than right action; it is right action in conditions of some measure of blessedness. I think it reasonable to assume that level of physical and social resources required for one to have the opportunity to achieve freely and autonomously chosen goals suffices for one to have the opportunity to lead a virtuous life. Likewise the resources required to develop the ability to deliberate well over ends—particularly exposure to an array of viewpoints and arguments, openly shared and debated within a pluralistic society—are the same resources required to develop the capacity for sound ethical deliberation. So we need not achieve anything more than the policy goals of Equal Liberty for each individual to have an equal opportunity at leading a flourishing life in the fullest sense. If the neo-eudaimonist is in want of a slogan, he may say that the good life is the life of Liberty, Achievement, and Virtue. In the remaining chapters, I take up the problems of the ground and the limits of the State's authority to create the conditions of Equal Liberty, and thus give each of its citizens an equal opportunity to lead lives of achievement and virtue.

References

Aristotle. *Eudemian ethics* [*Ethica Eudemia*] Oxford Classical Texts, ed. R.R. Walzer, and J.M. Mingay. Oxford: Oxford University Press.

Aristotle. *Nicomachean ethics* [*Ethica Nicomachea*] Oxford Classical Texts, ed. L. Bywater. Oxford: Oxford University Press.

Aristotle. *Politics* [*Politica*] Oxford Classical Texts, ed. W.D. Ross. Oxford: Oxford University Press.

Berker, S. 2007. Particular reasons. *Ethics* 118: 109–139.

Binmore, K. 2011. *Natural justice*. Oxford: Oxford University Press.

Boesch, C. 1994. Cooperative hunting in wild chimpanzees. *Animal Behavior* 48: 653–667.

Bowles, S., and H. Gintis. 2011. *The cooperative species: Human reciprocity and its evolution*. Princeton: Princeton University Press.

Brandom, R. 1994. *Making it explicit*. Cambridge, MA: Harvard University Press.

Broadie, S., and C. Rowe (eds. and trans.). 2002. *Aristotle: Nicomachean ethics: Translation, introduction, and commentary*. Oxford: Oxford University Press.

Burge, T. 2010. *Origins of objectivity*. Oxford: Oxford University Press.

Call, J. 2009. Contrasting the social cognition of human and nonhuman apes: The shared intentionality hypothesis. *Topics in Cognitive Science* 1: 368–379.

Call, J., and M. Tomasello. 2008. Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science* 12(5): 187–192.
Castelfranchi, C., and E. Lorini. 2003. Cognitive anatomy and functions of expectations. In *Proceedings of the First European Cognitive Science Conference (EuroCogSci 2003)*, ed. F, Schmalhofer, R.M. Young, and G. Katz, 377–389. Mahwah: Lawrence Erlbaum.
Dancy, J. 2003. The particularist's progress. In *Moral particularism*, ed. Brad Hooker, 130–157. Oxford: Oxford University Press.
Deacon, T. 1997. *The symbolic species: The co-evolution of language and the brain*. New York: WW Norton.
deVries, W. 2005. *Wilfrid Sellars*. Chesham: Acumen.
Doris, J. 2002. *Lack of character*. Cambridge: Cambridge University Press.
Gärdenfors, P., I. Brink, and M. Osvath. 2012 The tripod effect: The co-evolution of cooperation, cognition, and communication. In *The symbolic species evolved*, ed. T. Schilhab, F. Stjernfelt, and T. Deacon. Dordrecht: Springer.
Gaus, G. 2011. *The order of public reason*. Cambridge: Cambridge University Press.
Geuss, R. 2001. *History and illusion in politics*. Cambridge: Cambridge University Press.
Gintis, H. 2010. *The bounds of reason*. Princeton: Princeton University Press.
Hardie, W.F. 1977. Aristotle's doctrine that virtue is a 'mean'. In *Articles on Aristotle vol 2: Ethics and politics*, 33–46. New York: St. Martin's Press.
Harman, G. 1999. Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society* 99: 315–331.
Heath, J. 2001. *Rational choice and communicative action*. Cambridge, MA: MIT Press.
Heath, J. 2008. *Following the rules: Practical reasoning and deontic constraint*. Oxford: Oxford University Press.
Iacoboni, M. 2003. Understanding intentions through imitation. In *Taking action: Cognitive neuroscience perspectives on intentional acts*, 107–138. Cambridge, MA: MIT Press.
James, S. 2003. Rights as enforceable claims. *Proceedings of the Aristotelian Society New Series* 103: 133–147.
Jensen, K., J. Call, and M. Tomasello. 2007. Chimpanzees are rational maximizers in an ultimatum game. *Science* 318: 107–109.
Kramer, M., N. Simmonds, and H. Steiner. 2000. *A debate over rights: Philosophical inquiries*. Oxford: Oxford University Press.
Leibowitz, U.D. 2013. Particularism in Aristotle's *Nicomachean ethics. The Journal of Moral Philosophy* 10(2): 121–147.
Majumdar, D. 2004. An axiomatic characterization of Bayes' rule. *Mathematical Social Science* 47(3): 261–273.
Mill, J.S. 1863/2002. *Utilitarianism*, ed. G. Sher, 2nd ed. Indianapolis: Hackett.
Miller, F.D. 1995. *Nature, justice, and rights in Aristotle's politics*. Oxford: Clarendon Press.
Muller, M.N., and J.C. Mitani. 2005. Conflict and cooperation in wild chimpanzees. *Advances in the Study of Behavior* 35: 275–331.
O'Neill, O. 1996. *Toward justice and virtue: A constructive account of practical reasoning*. Cambridge: Cambridge University Press.
Peterson, C., and M. Seligman. 2004. *Character strengths and virtues: A handbook and classification*. Oxford: Oxford University Press.
Rakoczy, H., F. Warneken, and M. Tomasello. 2008. The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology* 44(3): 875–881.
Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.
Raz, J. 1994. Liberating duties. In *Ethics in the public domain*, 29–43. Oxford: Clarendon Press.
Reeve, C.D.C. 2013. *Aristotle on practical wisdom: Nicomachean ethics VI, Translated with an introduction, analysis, and commentary*. Cambridge, MA: Harvard University Press.
Sabini, J., and M. Silver. 2005. Lack of character? Situationism critiqued. *Ethics* 115: 535–562.
Scott, J.F. 1971. *Internalization of norms: A sociological theory of moral commitment*. Englewood Cliffs: Prentice Hall.

Sellars, W. 1956. Empiricism and the philosophy of mind. In *Minnesota studies in the philosophy of science*, vol. I, ed. H. Feigl and M. Scriven, 253–329. Minneapolis: University of Minnesota Press.

Sellars, W. 1957. Inference and meaning. In *Intentionality and the mental: Minnesota studies in the philosophy of science*, vol. II, ed. H. Feigl, M. Scriven, and G. Maxwell, 507–539. Minneapolis: University of Minnesota Press.

Sellars, W. 1968. *Science and metaphysics: Variations on Kantian themes*. London: Routledge & Kegan Paul.

Sellars, W. 1974. Meaning as functional classification. *Synthese* 27: 417–437; 423–424.

Sellars, W. 1981. Mental events. *Philosophical Studies* 39(4): 325–345.

Sen, A. 2007. Rational choice: Discipline, brand name, and substance. In *Rationality and commitment*, ed. F. Peters and H.B. Schmid, 353–354. Oxford: Oxford University Press.

Sherman, J. *Dialectical deadlock and the function of legal rights*. Unpublished MS.

Sherman, J. 2010. A new instrumental theory of rights. *Ethical Theory and Moral Practice* 13(2): 215–228.

Sinott-Armstrong, W. 2005. You ought to be ashamed of yourself (When you violate an imperfect moral obligation). *Philosophical Issues* 15(1): 193–208.

Snow, N. 2009. *Virtue as social intelligence: An empirically grounded theory*. New York: Routledge.

Stjernfelt, F. 2012. The evolution of semiotic self-control. In *The symbolic species evolved*, ed. T. Schilhab, F. Stjernfelt, and T. Deacon. Dordrecht: Springer.

Tomasello, M. 2001. *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tomasello, M. 2009. *Why we cooperate*. Cambridge, MA: MIT Press.

Tomasello, M. 2010. *Origins of human communication*. New York: Bradford.

Tomasello, M. 2014. *A natural history of human thinking*. Cambridge, MA: Harvard University Press.

Tomasello, M., and M. Carpenter. 2007. Shared intentionality. *Developmental Science* 10(1): 121–125.

Tomasello, M., and H. Moll. 2013. Why don't apes understand false beliefs? In *The development of social cognition*, ed. M. Banaji and S. Gelman. New York: Oxford University Press.

Tomasello, M., M. Carpenter, J. Call, T. Behne, and H. Moll. 2005. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* 28: 675–691.

Wittgenstein, L. 1953/2009. *Philosophical Investigations*, Trans. G.E.M. Anscombe, ed. P.M.S. Hacker, and J. Schulte. Oxford: Wiley-Blackwell.

Wormsley, W.E. 1993. *The white man will eat you! An anthropologist among the Imbonggu of New Guinea*. London: Harcourt Brace Jovanovich College Publishers.

Chapter 15
The Moral Justification of State Authority

1 Introduction

In order for the rights of individuals to be protected, coercive measures are sometimes required. The type of entity best suited to coercively enforcing rights—even by the lights of a staunch libertarian like Nozick—is the State. "The State" is, of course, an abstract concept. It is made up of individual persons—authorities—who have the power, in virtue of their position as representatives of the State, to act to enforce the rights of others. Raz's Normal Justification Thesis describes the conditions under which an authority's possession and exercise of this power is morally (as opposed to legally or politically) justified. We then say the authority is legitimate. But legitimate authorities do not merely have a morally justified power to enforce the rights of others on their behalf. They normally also have a moral right (again, in virtue of their position as representatives of the State) that those subject to their authority obey their directives. It is the authorities—natural persons—not the State itself—an abstract entity and artifical, legal person—who must possess these moral rights, if any exist. Artificial persons do not have moral rights. My ultimate goal in this chapter is to explain, in a way that satisfies the justification constraint, the conditions under which an authority has a moral right to compliance.

I endorse the meticulous argument of A. John Simmons, and condemn as hopeless all attempts other than the Contract Theory at establishing a content-independent political obligation—as opposed to a moral duty—to obey an authority (Simmons 1979). Since, as Simmons also argues, the Contract Theory is practically inapplicable, I focus on grounding a moral duty to obey the directives of a legitimate authority, and the right to compliance possessed by such an authority. Raz's Normal Justification Thesis does not address the issue of an authority's right to compliance with his directives. With my theory of rights as a foundation, I argue for a teleological justification of authority that explains why legitimate authorities have a right to compliance, thus yielding a fuller and more plausible account of legitimate authority.

Part of this chapter was previously published as Sherman (2010).

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_15

2 Rights and Authority

My teleological theory of rights explains why enforcing rights is justified. I now argue that it also succeeds in explaining why a legitimate authority, the appropriate agent of rights-enforcement in a political society, has a moral right of his own to compliance with his authoritative directives. Joseph Raz's Service Conception of Authority, which I introduced in Chap. 13, is one of the most influential contemporary accounts of the conditions under which the powerful exercise their power legitimately. Raz has recently clarified and refined various aspects of his view, and responded to many of the numerous objections to his position that have accrued over the years (Raz 2006). Problems do remain, however, which those of us who wish to defend the Service Conception, or something very close to it, must address. My goal is to resolve some of these problems by integrating the insights of the Service Conception into my broader framework for understanding moral rights and duties. After reviewing the content of the Service Conception (and expanding it somewhat), I introduce three problems which a complete account of legitimate authority must answer, and which are not addressed by Raz's view. By incorporating the main elements of the Service Conception into my account of moral rights and duties, I construct a theory of legitimate authority with the resources needed to provide solutions to the problems I introduce. I conclude by resolving each of the problems from the perspective of this new theory.

3 The Service Conception

A legitimate authority is one who has "a right to rule, where that is understood as correlated with an obligation to obey on the part of those subject to the authority" (Raz 1986, p. 23).[1] The Service Conception attempts to give an account of the conditions under which an authority's directives create duties for those subject to the authority to do what they have been directed to do. It is important to note that the Service Conception is concerned with the *moral* legitimacy of authority (as opposed to legal or political legitimacy). It construes a legitimate authority as a practical authority who imposes *moral* duties on those subject to his authority (Raz 1986, p. 53). A *de facto* authority's exercise of power is legitimate if, and only if, his directives succeed in creating such duties. Raz understands *de facto* authority as the power to exert direct influence over others' actions, combined with a claim that one's authority is legitimate—i.e. that one is entitled to exercise that power (Raz 1979, p. 7). Legitimate authority is then a species of *normative* power: a legitimate authority is a *de facto* authority who influences others' actions by creating new

[1] This idea, that a legitimate political authority has a claim-right that those subject to his authority comply with the directives which he issues through the powers of his office, can be traced back to Aristotle (Aristotle *Politics* III.13 1283b21-23, cited in Miller 1995, p. 107).

duties for them. A legitimate authority creates new duties by issuing authoritative directives. By creating these new duties, the authority changes the normative situation of those subject to his authority, rather than merely identifying the normative situation they are already in, or pressuring them to do what they already ought to do. The authority's directives thus *make a difference* to the normative situation of those subject to his authority, and it is in making such a difference that his exercise of normative power consists (Raz 1986, p. 48).

Recall that the Service Conception consists of three theses and a definition of 'duty'. The three theses are the Dependence Thesis (DT), the Normal Justification Thesis (NJT), and the Pre-emption Thesis (PT):

> DT: Authoritative directives should be based on the balance of relevant reasons that already independently apply to those subject to the directives (Raz 1986, p. 47).
>
> NJT: Authoritative directives should make those subject to the authority likely better to comply with the relevant, independently applying reasons by accepting and following the directives as authoritative, rather than by trying to follow the applicable reasons on their own. Demonstrating this is the normal way to justify an exercise of authority (Raz 1986, p. 53).
>
> PT: If DT and NJT are satisfied, then the fact that an authority has issued a directive is a reason to do what is directed which excludes and replaces (at least some of) the relevant, independently applying reasons (Raz 1986, p. 57).

These theses are then interpreted in the light of Raz's characterization of duties as pre-emptive reasons for action. Recall that a pre-emptive reason for action is a special type of protected reason. A protected reason is both (a) a first-order reason to perform (or refrain from performing) some action; and (b) a second-order exclusionary reason; that is, a reason not to act on other first-order reasons which compete with the first-order reason referred to in (a). Authoritative directives create new protected reasons for action. When the first-order component of a protected reason favors performing the act that is supported by the overall balance of pre-existing reasons (i.e., when the authoritative directive that creates the new reason satisfies the Dependence Thesis), we call it *conclusive*. A pre-emptive reason is a conclusive protected reason whose exclusionary component excludes all present competing first-order reasons. This is a duty. When an authoritative directive satisfies the Normal Justification Thesis, the exclusionary reason it creates does exclude all pre-existing reasons against doing what the directive orders (and replaces those pre-existing reasons for doing what the directive orders). Thus, directives that satisfy the DT and the NJT create new pre-emptive reasons—new duties. According to the Service Conception, therefore, a legitimate authority is a *de facto* authority whose directives create new pre-emptive reasons for action, and thus new duties, for those subject to the authority (Raz 1986, p. 60).

Raz's presentation of the Service Conception requires elaboration and expansion on a couple of points. The first is that there is no conflict between the claim that there are legitimate authoritative directives and the claim that the good and responsible moral agent deliberates about what he ought to do. Authoritative directives do not cut off deliberation. Rather, they *change the normative landscape*; they change what there is for the good and responsible moral agent to use as input for deliberation. When a legitimate authoritative directive to perform some particular action has

been issued, there is simply nothing for the process of deliberation to consist in other than the recognition of that directive as creating a pre-emptive reason for action. When the pronouncement of a legitimate authority is not a particular directive but rather a law, it falls to the individual moral agent, perhaps with the help of a history of judicial precedent in which that law is interpreted, to determine what is required of him with respect to a choice that he faces. Directives, moreover, need not only concern the actions among which an agent must make a choice. They may instead concern the deliberative act itself. The paradigmatic example of this is a judge instructing a jury to disregard some statement of a witness, claimant, defendant or prosecutor. Such an instruction creates a duty to come to a conclusion about the facts at issue in the case without factoring the offending statement into the deliberations which will lead to that conclusion. The responsibility for deliberating—for weighing all the evidence which is admissable—and the freedom to reach the conclusion which (as far as they are able to determine) the admissable evidence supports, still rests with the members of the jury.

The second point is that the pronouncements of legitimate authorites are not limited to directives and laws that create duties. They also create *permissions*. An authority who issues a permission either to ϕ or to ψ aims to create a conclusive reason for the addressee to act as if the reasons that favor ϕ-ing are exactly as strong as the reasons that favor ψ-ing, and to disregard any reasons for believing that (or acting as if) this is not the case. In the spirit of the Service Conception, we should say that such an authoritative permission-creating act is legitimate—and so such a reason is created—just in case the reasons are equally balanced, and the permission makes the agent more likely to conclude that he should simply make whatever choice he is inclined to make, than he would be if he were to try to determine the balance of reasons on his own. Permissions, like directives, may concern the deliberative act itself. One may, for example receive a permission to treat two pieces of evidence as equally important.

Raz does discuss permissions outside of the context of his theory of authority. He argues that supererogatory action is possible because, in situations where the balance of reasons in fact favors an action that requires a great sacrifice on the part of the agent, there exists a permission to act as if the reasons favoring an action requirirng less sacrifice were equally strong. To perform the actions which is actually favored by the balance of reasons is then supererogatory (Raz 1975). I have no use for such a notion of permission. First, it is not at all clear what sort of fact could ground such a natural permission (i.e. one not created by an authority). As Jonathan Dancy has pointed out in his critique of Raz's theory of supererogation, the only likely candidate is a fact like "this is very demanding." And if that is the sort of fact that grounds these permissions, supererogation cannot be a matter of degree—it kicks in only when the going gets really tough, at which point we are permitted to ignore the call of reason and do as we like (Dancy 1993). This is a terribly unattractive view of supererogation, and I have already provided an outline of a much more satisfactory account—one which does not require any notion of permission—in Chap. 13. Reason permits us to choose however we are inclined when the reasons supporting our options are equally strong (though we remain morally free—in the

sense of not deserving blame—to choose what is less supported, provided that its reasons have not been excluded). This is the only notion of natural permission we need; and the legitimacy of an authoritative permission depends on there being such a natural permission.

I offer these remarks simply by way of filling out the Service Conception, so that our understanding of it is as complete as it can be before we turn a critical eye toward it and expose some of the defects which must be remedied. The problems we shall explore in this chapter all concern authoritative directives to perform some action.

4 Edmundson on Legitimacy

William Edmudson has made a powerful argument against the claim that being a legitimate authority entails that one's authoritative directives create enforceable duties of compliance (and thus create a right to compliance). He calls this the "Strong Legitimacy Thesis," and contrasts it with the "Modest Legitimacy Thesis," according to which legitimate directives only create an enforceable duty not to interfere with their forceful administration (Edmundson 1998, p. 43). Edmundson, like Raz, believes that being a *de facto* authority entails *claiming* to create enforceable duties of obedience—a proposition which Edmundson calls the "Strong Authority Thesis." Edmundson nonetheless thinks that only the Modest Legitimacy Thesis can be defended, because he does not accept what he calls "The Warranty Thesis":

> If being an *X* entails claiming to ϕ, then being a legitimate *X* entails *truly* claiming to ϕ. (Edmundson 1998, p. 43)

He argues that the Strong Authority Thesis requires the Warranty Thesis, but that the Warranty Thesis is false. In place of the Warranty Thesis, he advocates what he calls "The Proximity Thesis":

> If being an *X* entails claiming to ϕ, then being a legitimate *X* entails *sincerely* claiming to ϕ, and coming within tolerable limits of doing so. (Edmundson 1998, p. 47)

Edmundson goes on to argue that the Proximity Thesis can only support the Modest Legitimacy Thesis, but that the latter is all the legitimacy we need (Edmundson 1998, ch. 3).

I doubt, however, that Edmundson succeeds in showing that we must reject any version of the Warranty Thesis. His argument runs as follows (Edmundson 1998, pp. 44–47). On one reading, the Warranty Thesis is obviously false as a claim about both theoretical and practical authority. On another reading, it is true as a claim about both theoretical authority and practical authority, but this reading is incompatible with the Strong Legitimacy Thesis. Let us begin with a particular species of theoretical authority: scientific authority. The first reading takes "ϕ" as something like "know exactly how some part of the world really works." On that reading, both

the Strong Authority Thesis and the Warranty Thesis are false. Scientific authorities need not claim to know how the world really works, and one need not know how the world really works in order to be a legitimate scientific authority. In the case of practical authority, we would take "ϕ" as something like, "know exactly what the one subject to the authority ought to do, in the objective sense of 'ought.'" Edmundson recognizes that no advocate of the Strong Legitimacy Thesis would want to make so strong a claim. Given my Peircean view of normative truth, I certainly would not.

He returns to the case of scientific authority, and considers taking "ϕ" as something like "know what theory of the way some part of the world works is most supported by the current balance of the best available evidence." And so we may likewise, in the case of practical authority, take "ϕ" to be something like "know what normative conclusion is most supported by the best available evidence." The Warranty Thesis now seems to be an acceptable claim, in both the theoretical and practical cases. Edmundson, however, sees acceptance of this reading of the Warranty Thesis as a fatal move for the proponent of the Strong Authority Thesis: "Defending the Warranty Thesis by toning down the claims made by theoretical authority makes it incongruous to insist on the Strong Authority Thesis" (Edmundson 1998, p. 46). But this simply is not so. There is no tension between these two claims.

The reason is that *de facto* authorities make not one claim, but *two*. They claim *both* that their directives create enforceable duties of compliance, *and* that they know what normative conclusion is most supported by the best available evidence. In other words, they claim that their directives create *both* a pre-emptive reason for *belief*, *and* a pre-emptive reason for *action*. This is the crucial way in which theoretical authority differs from practical authority. To say that a scientific authority's pronouncements are pre-emptive reasons for belief is to say that the fact that a given scientific authority tells one that the world works thusly is both a first-order reason to believe that the world works thusly, and an exclusionary reason to disregard other reasons (or what appear to be other reasons) to believe otherwise. To accept the scientist's authority is to adopt the relevant belief *for the reason that* the scientist advocates it. A scientific authority is legitimate, then, just in case one is more likely to believe what one ought to believe about the world—in the intersubjective, best-available-evidence sense of "ought"—than one would be if one endeavored to evaluate the evidence for oneself. All the same is true of other types of theoretical authority, which are closer to practical authority in their subject matter. Legitimate *prudential* authority is a source of pre-emptive reasons for conditional normative beliefs. The investment advisor, for example, who tells one to diversify one's portfolio if one wants to minimize one's risk, is acting as a prudential authority. We may also distinguish a theoretical species of legitimate *moral* authority, which is a source of pre-emptive reasons for unconditional normative beliefs. An example might be the non-binding arbitrator who determines what specific resolution to a conflict would produce the fairest outcome for all parties involved. Authorities of these types create pre-emptive reasons for belief analogous to those created by scientific authorities, and have analogous legitimacy conditions.

Genuinely practical legitimate authority, on the other hand—political and legal authority— is a source of both pre-emptive reasons for unconditional normative beliefs *and* pre-emptive reasons for action.[2] The Warranty Thesis concerns the former; the Strong Authority Thesis concerns the latter. There is no tension between the two. Edmundson almost manages to find his way to this conclusion himself. He recognizes that neither theoretical authorities nor purely prudential practical authorities create *duties*—i.e., on my and Raz's view, pre-emptive reasons for action (Edmundson 1998, p. 46). He sees that there is some sense in which political authority is more deeply normative than the other types (Edmundson 1998, p. 47). But he cannot see the full significance of the disanalogy. Note, finally, that there is no tension between the reading of the Warranty Thesis that I have adopted, and the Strong Legitimacy Thesis. In order for the directive of a legitimate practical authority to create a pre-emptive reason for action of the right type—that is, an enforceable moral duty of compliance—it must be the case that the one subject to the authority is less likely to end up acting as he ought—again, in the intersubjective sense—by examining the independently applying reasons himself, than he would be by accepting the directive. It need not be the case that the authority knows what the subject ought to do, in the objective sense of "ought."

5 Three Problems with the Service Conception

I think the Service Conception is correct, so far as it goes. I will not, therefore, attempt to add to Raz's recent defense of it against the objections it has already faced. My concern, rather, is with certain problems that any complete account of legitimate authority must resolve, and which Raz's account does not address. I now introduce three such problems.

5.1 Legitimate Authority and Enforceable Duties

The first problem facing Raz's view concerns the nature of the duties that are imposed by legitimate authorities. Practical authorities, at least within politically advanced societies, act within legal systems. Their authoritative directives impose legal duties and confer legal rights correlative with these duties. Fulfillment of these

[2] In fact, the situation is rather more complicated than this. The directive of a legitimate practical authority creates a pre-emptive reason to believe that, before the directive was issued, the directed action was right. But the pre-emptive reason for action it creates is also a reason for belief—just as a natural moral duty is. The latter is a reason to believe that other actions are wrong—which may not have been the case before the directive was issued. We respond to both the former and the latter by passing over competing reasons in deliberating about what we ought to do. But only the latter makes it morally inappropriate to attend to these competing reasons in deliberation, because only the latter excludes them as reasons for action; the former only excludes them as reasons for belief.

duties can be exacted, coercively if necessary, by practical authorities through the use of legally conferred powers. But when we seek to determine whether an authority is legitimate, we are not concerned with the question of whether his authoritative directives succeed in imposing legal duties according to the legal system within which he acts. We are concerned instead with whether his directives create moral duties. The point that the Service Conception overlooks in attempting to specify conditions under which moral duties are created is the fact that unlike legal duties, moral duties are not necessarily correlative with claim-rights, and are not necessarily justifiably enforceable through the exercise of any power. To explain how authorities create moral duties, then, is not necessarily to explain how they create justifiably enforceable moral duties or confer moral rights.

We have already seen that unlike moral rights, there are many moral duties that cannot be justifiably enforced. We do not simply understand legitimate practical authorities as authorities who create moral duties. We understand them as authorities who create justifiably enforceable moral duties. When a duty generated by a legitimate directive is violated, it is morally justified that the violator be compelled, coercively if necessary, either to fulfill the duty or to give compensation for having violated it.

The Service Conception, however, merely specifies the conditions under which an authority's directive creates a moral duty for the one subject to the directive. It fails to make clear why such a duty is justifiably enforceable. If part of being a legitimate practical authority is having the power to impose justifiably enforceable duties, as it surely is, we should be able to explain what makes this the case. Such an explanation will have two parts. It will first distinguish between those duties that can be justifiably enforced and those that cannot, by articulating relevant differences in the pre-emptive reasons for action that constitute duties of these types. It will then demonstrate that directives issued by legitimate authorities ground justifiably enforceable duties. The Service Conception does neither. The view I will develop succeeds in doing both. In addition, legitimate authorities are normally morally justified in enforcing the duties which they have imposed, coercively if necessary. A complete account of legitimate authority must explain why they are justified in doing so; and as Raz acknowledges, the Service Conception does not attempt to do this (Raz 2006, p. 1037). My account will address this issue as well.

5.2 Practical and Theoretical Authority

The second problem concerns the difference between practical and theoretical authorities. A theoretical authority is an expert. The proclamations of theoretical authorities on matters within their areas of expertise are pre-emptive reasons to believe the contents of those proclamations. Suppose, for example, that an international team of climate scientists proclaims that the current rate of global warming

will lead to a rise in the sea level of 18–59 cm by the year 2100.[3] Climate science is an extraordinarily complicated field, and a great deal of specialized training is required for one to correctly gather and interpret the evidence that bears on the rate of climate change and the most likely effects of that change. Even an intelligent, well-educated non-climate scientist would have a much better chance of holding true beliefs about climate change by accepting the proclamations of the experts as authoritative. To do so is not only to take their status as experts as a reason to believe what they say; it is also to take it as a reason not to weigh the evidence on either side of the issue for oneself. Recognizing them as authorities is in part a matter of recognizing that one's own attempts to interpret and weigh all the relevant evidence would be much less likely to succeed—would instead likely result in error and confusion. The fact that the experts have reached a consensus *replaces* the independent reasons on which that consensus is founded as the basis for the average person's belief in the experts' conclusions.

Theoretical authorities do more than report their findings to the public; they also advise on the best course of action to achieve goals related to their areas of expertise. Climate scientists might determine, for example, that the most effective way for the average American to combat global warming is to stop taking his car on trips that measure less than 2 km in total. They might then advise that all Americans do this as a way to advance efforts to combat global warming. Raz observes that we do not consider merely theoretical authorities to have any authority *over* us (Raz 1986, p. 63). Though they may advise, they are not entitled to command. They may provide us with pre-emptive reasons for belief, and it may be true that if what we want is to achieve the goals to which their expertise is relevant, we ought to accept what they say as authoritative. But they do not provide us with pre-emptive reasons for action, as practical authorities do. They do not impose duties on us. I think all this is perfectly true, but a complete account of legitimate practical authority must explain why only practical authorities are capable of imposing duties.

Raz's account fails to give such an explanation. To see this, let us consider the following hypothetical example. Suppose the chairman of the Intergovernmental Panel on Climate Change appears on television, orders the American people to stop taking their cars on trips that total less than 2 km, and claims he is entitled to give this order. We should say, as Raz no doubt would, that his order is impotent; it does not impose on the American people a duty to refrain from driving very short trips. But why so?[4] Just as the Dependence Thesis seems to require, the order he has given is (let us say) based on the balance of relevant, independently applying reasons. And just as the Normal Justification Thesis seems to require, those who are ordered are likely better to comply with the balance of relevant independently applying reasons by accepting the order as authoritative and following it, than they would be if they attempted to follow those reasons on their own. Why then should the Pre-emption

[3] As, in fact, they have (IPCC 2007, p. 747 *et seq.*).

[4] Alternatively, we might claim that this could not have been a genuine order at all, that the person in question had merely aped the speech-act of giving an order. But even if we do say this, the question still stands.

Thesis not hold in this case, and a duty thus be imposed on those to whom the order is addressed?

Raz does have a ready response to this concern (Raz 2006, pp. 1032–1037). Insofar as the Dependence and Normal Justification Theses concern *de facto* practical authorities, they set up the conditions under which such authorities exercise their power legitimately. These theses are meant to specify the conditions under which practical authorities create moral duties of compliance. These theses do also specify the requirements for being a theoretical authority. But merely theoretical authorities are distinguished from practical authorities precisely by the fact that the former do not have the power to directly influence the actions of others, and cannot create duties for others.[5] This response, however, defers the crucial question, rather than answering it. It is true that a merely theoretical authority does not possess the power to directly influence the actions of others. What must be explained is why power should be what accounts for the practical authority's ability to impose moral duties on others through his directives, versus the theoretical authority's lack of this ability. In the example above, the climate scientist claims to be entitled to exercise authority over others, and succeeds in issuing a directive which, if accepted and followed as authoritative, would lead its addressees to better comply with the relevant, independently applying reasons. All he lacks is practical, effective power over them. It is not immediately clear why the possession of power should make this kind of normative difference—should enable the powerful to create duties, something which the expert cannot do. The second problem facing the Service Conception is thus to provide a plausible explanation of this difference.[6]

5.3 Pre-existing Moral Duties

The third problem concerns a thesis which the Service Conception aims to undermine: what Raz calls the No Difference Thesis ((Raz 1986, p. 48). This thesis states that the directives of legitimate authorities make no difference to the normative situations of their addressees. A legitimate directive simply tells its addressee to do what he ought to do anyway. The Service Conception counters this by claiming that legitimate directives replace the independent reasons for whatever action they direct, and furthermore, that they replace these first-order reasons with a reason of a different kind—a pre-emptive one. Authoritative directives thus "turn 'oughts' into

[5] This is also Raz's response to what he calls the 'qualification objection': the charge that the Service Conception does not articulate what makes someone a legitimate authority, but rather what characteristics would make someone a good authority. An error closely related to the one that underlies the qualification objection is made by Stephen Darwall in his recent critique of the Service Conception. Darwall understands the Service Conception as entailing the claim that merely theoretical authorities can create duties for others (Darwall 2009).

[6] This objection to the Service Conception has recently been made by Jonathan Quong (Quong 2011, pp. 114–115). As we will see at the end of this chapter, Quong is quite mistaken in thinking that the problem cannot be solved.

duties," and in this way are supposed to make a significant difference to the normative situation of their addressees (Raz 1986, p. 60).

Raz has recently responded to a challenge to the Service Conception which focuses on the existence of moral reasons which are independent of authoritative directives (Raz 2006, p. 1022). Independent of any authoritative directive, it is the case that we ought not to murder other people. Furthermore, it seems implausible to assert that the fact that we are directed not to murder others replaces our independent reasons for not murdering them. The best reason for not murdering others is not the fact that the law so directs; it is the value of human life, or some similar consideration. Raz acknowledges that this point is perfectly correct, and asserts that it is no problem for the Service Conception to admit that individuals should sometimes be guided by independent reasons, rather than by the directives of authorities. In these cases, he claims, the role of the directive is merely to ground an exclusionary reason—a reason to ignore any reasons there may be which conflict with the reason we ought to act on (Raz 2006, pp. 1022–1023). By serving this exclusionary function, the directive turns an 'ought' into a duty.

It is not merely the case, however, that one ought not, on balance, murder other people. There is a moral duty not to do so; and one has this duty independent of any authoritative directive. So it cannot be the case that when the law directs one not to murder others, it is turning an 'ought' into a duty. Raz's response, then, seems to imply that in cases of pre-existing moral duties, the only normative difference made by an authoritative directive is to place the addressee under a second, redundant, less normatively significant duty. Assuming that a second duty to perform (or refrain from) the action in question really is imposed, then perhaps this response still manages to reject the No Difference Thesis. But this is certainly not the kind of normative difference that Raz originally set out to establish. We might say that, at least in cases of pre-existing moral duty, the response brings the Service Conception in line with a Not Much Difference Thesis: authoritative directives only serve to generate redundant, relatively insignificant moral duties. This clearly runs counter to the spirit of the Service Conception and the idea that acts of legitimate authority make a real normative difference. The view of legitimate authority I will develop succeeds in establishing that even in cases of pre-existing moral duty, the normative difference made by the exercise of legitimate authority is significant.

6 Incorporating the Service Conception

In this section I lay out an account of legitimate practical authority that incorporates the insights of the Service Conception into my general framework of moral duty. To facilitate the development of my account, I will sketch some simple examples of a practical authority exercising his power, and then fill in the account of legitimacy with reference to these examples. Suppose that Solomon is the king of a small

nation, whose population consists of only a couple of tribes.[7] Solomon is the supreme practical authority within his kingdom. He makes the laws of the kingdom by issuing proclamations; he serves as magistrate when disputes arise between his subjects; and he commands the kingdom's security forces. All official acts, regardless of who carries them out, are done in his name.[8]

Solomon's kingdom is under threat of attack by a neighboring state. An accomplished military strategist, he has examined the possible paths of approach and determined that the enemy will most likely attack from the southwest. He orders Joab, one of his field marshals, to occupy a hill on the southwestern border of the kingdom, so that his forces will have the higher ground in battle. We are interested in the question of whether Solomon has legitimate authority over Joab. We should say that Solomon does have legitimate authority over Joab if the fact that he has given Joab an order is (1) a pre-emptive moral reason for Joab to perform the action that he has ordered Joab to perform; and (2) sufficient to justify coercive action by Solomon if Joab refuses to carry out the order. For this would mean that by commanding Joab, Solomon places him under an enforceable moral duty to comply with the order, and that Solomon has a moral right to his compliance with the order, and would be justified in obtaining that compliance coercively.

According to the account of moral duty sketched in the previous chapter, we should determine whether Joab owes Solomon an enforceable moral duty by examining the nature of Solomon's interests. Solomon is a *de facto* practical authority who has issued a command to one over whom he claims authority, and within an area which he claims falls within his authority. As such, he has an interest in compliance from the one whom he has ordered. Our question is: is this interest sufficient to ground a justifiably enforceable duty for Joab? The account of moral duty just outlined provides a precise guide for determining the answer to this question. We first ask whether Solomon's interest in compliance should be classified as an important one. As the supreme practical authority within his kingdom, commanding the security forces is one of the primary functions Solomon performs. Since leading his kingdom is the primary project he is engaged in, and serving as military commander is one of the major aspects of that project, Solomon cannot be said to succeed in his primary project without serving as an effective commander of the military, and doing all that he can to preserve his kingdom's security. Since he cannot be an effective military commander if his subordinates do not comply with his orders, there is a strong *prima facie* case for classifying his interest in their compliance as an important one.

In order to determine definitively whether Solomon has an important interest in compliance, however, we must incorporate one of the insights of the Service

[7] The example is not meant to be biblically accurate.

[8] The point of the example is to avoid complications resulting from features of more sophisticated political systems, such as separated powers and representative government. I do believe the account of legitimate authority offered applies to these more complex systems. But the additional difficulties of describing and analyzing authorities and authoritative acts in those contexts would only obscure the presentation of the basic view.

Conception. If Solomon were unfit to command, owing to ineptitude in military matters, compliance with his orders would not further his overall project of being a successful leader. Such compliance might, rather, result in defeat and serious harm to his kingdom. We will only conclude, then, that he has an important interest in exercising his *de facto* authority if his use of that authority satisfies the Dependence Thesis. The authoritative directives that he issues must depend on and reflect the relevant independently applicable first-order reasons.

That Joab is relevant to the satisfaction of this interest is clear. The order was given to him. Let us proceed, then, to the question of whether the satisfaction of Solomon's interest in compliance is necessary to the flourishing of his kingdom. First we must ask whether the arena within which authority is being exercised is one which is of great significance to the welfare of the society. Military command certainly meets this condition, as do many of the arenas within which authority has traditionally operated, and continues to operate, in most societies. So given that we are assuming Solomon exercises his authority in accordance with the Dependence Thesis, we have good grounds for asserting that in exercising his authority he will *contribute* to the common good. But according to the account of moral duty laid out in the previous section, his order will only ground a duty of compliance if it is *necessary* to the common good that he exercise his authority effectively. Here we must incorporate another of the Service Conception's basic insights. If it were the case that Joab would be just as likely to make the right military decision if left to his own devices, then Solomon's exercise of authority would not be necessary. However, if Joab would be more likely to perform the correct action by accepting Solomon's order as authoritative and acting on it than he would be if he were to consider the relevant reasons himself, then we have good reason to say that Solomon's exercise of authority is (practically) necessary to the common good. And this is just to say that as an authority, Solomon satisfies the conditions of the Normal Justification Thesis. As an authority who satisfies the NJT, therefore, Solomon's interest in compliance is sufficient to ground a duty of compliance for his subjects.

For this duty to be enforceable, the interest must be of such central significance to social welfare that the cost of allowing it to go unsatisfied outweighs any costs associated with enforcing the duty. Again, in the case of military command and many of the other arenas in which authority has traditionally operated, this condition is satisfied. The duty of compliance with Solomon's order is enforceable, and Solomon therefore possesses a right to compliance correlative with this duty. And in virtue of his position within the military, Solomon's interest in compliance grounds sufficient justifying reason for him himself to obtain compliance coercively, should that be necessary. As *de facto* authority, he is capable of doing so with minimum disruption to social order. *Qua* military commander, therefore, Solomon is a legitimate practical authority.

This is the basic account of legitimate authority I have to offer on basis of my general account of moral duty. But the basic account does not succeed in explaining the nature of all exercises of legitimate authority. A slightly different account, still based in the incorporation of the Service Conception into my basic framework for moral duty, is needed to accommodate one of the central ways in which legitimate

adjudicative authority is exercised. To see this, let us consider another example. Solomon is the magistrate within his kingdom. He settles disputes between his subjects, and does so according to the laws which he himself has proclaimed. Because he values the rule of law, he considers himself to be bound by these laws in his capacity as magistrate. Suppose that two of his subjects, Cain and Abel, appear before him to argue a claim. Abel claims that Cain has failed to fulfill a contract, and Cain disputes this, maintaining that he has fulfilled their contract's terms. Solomon hears the evidence offered by both sides, and rules in Abel's favor. He orders Cain to compensate Abel according to the law governing defaults on contracts.

In delivering his judgment, Solomon is exercising his practical authority as magistrate. Since we want to know whether he is exercising that authority legitimately, we need to know whether Cain has a moral duty to comply. But there are important disanalogies between this case and the case of Solomon and Joab. If Solomon's authority is legitimate, then the fact that Solomon has delivered this verdict makes it the case that Cain now has a moral duty to provide Abel with compensation. But to whom is this duty owed? From a contemporary legal perspective, we would say that Cain owes a legal duty to Abel, and that Abel has a claim-right against Cain that this duty be fulfilled. If Cain fails to compensate Abel, Abel would be at liberty to request that his claim-right be enforced by the court. The court would then have a duty to Abel to do so by exercising its power of enforcement. However—and here is the important point for our purposes—Cain would not have a legal duty to the court to comply with the ruling, since the court would have no claim-right to compliance against Cain. If Cain should fail to provide compensation to Abel, his government would not be free to exact compliance from him on its own. From a moral perspective, this is as it should be. It is plausible to claim that the state would be wrong to interfere in a private dispute in this way. The responsibility, moral and legal, for taking action against Cain lies with Abel in this case.

Assuming again a moral perspective on our example, it seems we should claim that Solomon's authority is legitimate if the following are true: (1) The fact that Solomon has issued a ruling is a moral pre-emptive reason for Cain to comply with that ruling—i.e. the ruling places Cain under a moral duty of compliance to Solomon; (2) once the ruling has been issued, Cain has a justifiably enforceable moral duty to compensate Abel (who has a moral right against Cain correlative with this duty); and (3) if Cain should fail to comply and Abel should request that his right be enforced, Solomon would be morally justified in enforcing Abel's right against Cain (in fact, he would have a moral duty to Abel to do so). Cain does not, however, owe an enforceable moral duty to Solomon; Solomon himself does not have a right to compliance against Cain.

Can my theory of moral duty serve as a guide for constructing an account of legitimate practical authority in this case as well? The first point to note is that my theory does not wrongly entail that Solomon has a moral right that Cain comply with his ruling. Let us assume that as a magistrate Solomon exercises his authority in accordance with the Dependence and Normal Justification Theses. The judgments he delivers are morally sound—they reflect the balance of first-order moral reasons. By complying with these judgments, moreover, those to whom they are

addressed are significantly more likely to do what they ought to do, than if they attempted to determine what they ought to do on their own. Solomon has a refined sense of fairness, and an excellent capacity for evaluating evidence and working out the normative implications of the facts of the cases he adjudicates. If this is so, then the fact that Solomon has ruled against him *is* a pre-emptive moral reason for Cain to comply with that ruling; the ruling *does* place Cain under a moral duty *to* Solomon. Solomon's interest in compliance is sufficient to ground that duty. But his interest does *not* ground a justifiably enforceable moral duty in this case. The reason for this is that if the winning side of an adjudicated dispute decides against requesting that their right be enforced, very little social good would result from the state enforcing the ruling anyway, on the basis of its interest in obtaining compliance with the judgments of the courts. And whatever good might result would certainly be outweighed by the social harm of a living under a state which indulged in that level of interference in private affairs. My theory, therefore, can account for Solomon's lack of a moral right to compliance in this sort of case.

Nonetheless, if Solomon is acting as a legitimate authority, then it must be the case that Cain is now under a justifiably enforceable duty to compensate Abel, and Solomon must be morally justified in enforcing Abel's right should Abel request that he do so. If Solomon's authority is legitimate, then, his ruling must in some way contribute to the formation of a justifiably enforceable duty owed by Cain to Abel. Can my account of legitimate authority explain how this happens? To begin, the account can certainly establish Cain's moral duty to compensate Abel. If as a matter of fact Cain did violate their agreement, then Abel has an interest in being compensated by Cain which is sufficient to ground a moral duty. It is of prime importance both to the parties involved and to society as a whole that contracts be honored, and that when they are not honored, that appropriate compensation be rendered.

Cain thus has a moral duty to compensate Abel before any authority intervenes in their dispute.[9] But would Abel be justified in compelling Cain to fulfill this duty? At this point the existence of a practical authority, and the quality of that authority, becomes relevant. If there is a practical authority to whom Abel could bring his complaint, and this authority exercises his power in a way that satisfies the Dependence and Normal Justification Theses, then it would certainly be better, both for the parties involved and for society as a whole, for Abel to bring his complaint to this authority, rather than seek satisfaction on his own. In fact, given that such an authority exists, the social harm that Abel would cause by enforcing the contract on his own would outweigh the good of having the contract enforced. By taking matters into his own hands, Abel would strike a blow against the integrity of the important social institution of the magistracy. So we should conclude that given that a practical authority like Solomon exists, Cain's duty to Abel is not enforceable *by* Abel.

[9] Cain may not believe that he is under any such duty before his case is heard, and if the case is a complicated one, it may be reasonable of him not to believe this. If so, we should still say that he is under a duty to Abel, because the simple fact is that he is. But we should also say that he is not blameworthy for failing to fulfill his duty prior to his case being decided.

But for a duty to be enforceable it need only be the case that there is *someone* who would be morally justified in securing its fulfillment coercively. So far I have given reason for believing that it would be *preferable* for Solomon to enforce Cain's duty, rather than for Abel to do so. Our next question is whether Solomon is really morally justified in doing this, and in particular, whether my account of legitimate authority can explain why this is so. Abel's interest in receiving compensation certainly grounds some reason for Solomon, *qua* adjudicative authority, to exact that compensation on his behalf. We should deny that Solomon is morally justified in enforcing Cain's duty to Abel if, and only if, there is some non-excluded reason which outweighs Solomon's reason to enforce the duty.

We can assert that Cain's interest in not being coerced fails to ground such a reason, since he is under a duty to compensate Abel. To have a morally relevant interest in not being coerced into doing something, one must be morally free to refrain from doing it. We have already seen that Abel would do more harm than good if he were to enforce the duty himself instead of having Solomon do it. In reaching that conclusion, we assumed that Solomon the magistrate was a *de facto* authority who satisfied the Dependence and Normal Justification Theses. It seems now that all we have left to inquire is whether, given these assumptions, the social good of Solomon's coercive enforcing of Cain's duty will outweigh the social harm. If it does not, then there would be a strong reason for Solomon not to enforce the duty either. But it is clear that in this case, the good does outweigh any harm. The worries which applied to the case of Abel enforcing the duty for himself do not apply here. Because Solomon is a *de facto* authority, his act of enforcement would contribute to social stability, rather than weaken it. And since Solomon exercises his power in accordance with the Dependence and Normal Justification Theses, no social harm could result from his coercive act of enforcement that would outweigh the good of his securing just compensation for Abel in an orderly fashion. Solomon alone, *qua* magistrate, is therefore morally justified in enforcing Cain's duty to Abel.[10]

Let us now assemble these individual conclusions into one coherent picture. Abel's interest in compensation grounds a moral duty for Cain, but one which is not initially enforceable. Once an authority like Solomon has determined that Cain has violated the contract, Cain is placed under a second moral duty *to* Solomon to compensate Abel. This duty is not enforceable either—Solomon has no right against Cain. But should Cain fail to compensate Abel, and should Abel file another complaint with Solomon, Solomon would be morally justified in enforcing Cain's duty to Abel. This implies that once Solomon has delivered his ruling, Cain's duty to Abel becomes an enforceable one—there is now someone who would be morally justified in coercively securing the duty's fulfillment. And this further implies that Abel is now in possession of a right to compensation against Cain. The fact that

[10] As I indicated earlier, the reason for Solomon to enforce Cain's duty is a pre-emptive one—Solomon has a duty to Abel to enforce Abel's right. Though the argument for this claim does not require any material which would be unfamiliar at this point, my present purposes do not require that I articulate it.

Solomon is a *de facto* practical authority who satisfies the Dependence and Normal Justification Theses, then, makes it the case that (1) his rulings create moral duties of compliance that are owed to him; (2) his rulings transform duties owed in private disputes into justifiably enforceable duties (and thus confer correlative rights to compensation); and (3) he is morally justified in coercively enforcing those rights on behalf of the right-holders. These are precisely the conditions Solomon had to meet in order to qualify as a legitimate authority. By incorporating the insights of the Service Conception into my general framework for moral duty, my account succeeds in uncovering the moral ground of the authority to resolve private disputes and to enforce those resolutions.

In the example, I have assumed that a correct application of the law to the case at hand will yield a decision that accords with the balance of independent moral reasons. However, the drafting of all laws is constrained by limitations of time, knowledge, and foresight, as well as by the necessity of dealing in generalities. And so even good laws—which is to say, laws which satisfy the Service Conception by being such that, when correctly applied, they normally direct one to do what is supported by the balance of independent reasons, and one normally is more likely to do this by following the law than by deliberating independently—can on occasion go obviously wrong. Correctly applying a legal rule is a matter of correctly interpreting it, and I agree with Raz that correct interpretation of the text of a legal rule depends on the intentions of its author—which need not be a natural person, but may be an institution (Raz 2009a). More specifically, I hold that the meaning of a legal rule is determined by the actual realized communicative intentions of its (institutional) author - those communicative intentions for which there is some evidence in the text produced. We might call this theory of legal interpretation a Moderate Actual Institutional Authorial Intentionalism (MAIAI).[11] However, I also agree with Raz that when the correct application of even a good legal rule yields an obviously morally wrong result, a judge ought to decide the case in accordance with the moral reasons.

I see such instances as calling for judges to exercise of the Aristotelian virtue of *epieikeia*—to correct the law when it goes astray (Aristotle *NE* V.1137a32-1138a3).[12] My own, neo-Aristotelian view of *epieikeia* is that it is the distinctively and characteristically judicial virtue, insofar as it is not merely excellent moral reasoning applied in the context of a legal case, but also satisfies a set of additional normative requirements which exist in virtue of the existence of a modern legal system.[13] A judge not only issues a judgment, but delivers an opinion justifying that judgment; and this opinion forms the basis for a precedent which influences future decisions. One part of *epieikeia*, then, is producing a justification for the morally right decision which is appropriately narrow or broad, based on a judgment of how

[11] My views on this topic have been extensively shaped by A.P. Martinich.

[12] The term is usually translated as "equity", but since this invites confusion with the concept of equity in the English common law—to which it does bear some limited similarity— I leave it untranslated.

[13] The question of how best to interpret Aristotle's own account of this notion is a complicated one, and pursuing it would serve no purpose here.

likely the law is to go astray in future more or less similar cases. A judicial decision is also enforceable. Having determined with whom fault lies and the nature of the appropriate punishment or damages according to the independent moral reasons, the judge must set the precise severity of punishment or quantitative amount of damages. In so doing, he must not exceed what can be morally justifiably exacted from the relevant party. Finally, as Raz has argued, the judge must justify his decision in the form of a novel interpretation of the text of the legal rule being corrected—novel in that the interpretation does not follow from the MAIAI model. Raz gives a number of good reasons for valuing this exercise. These include the need to adjust the law itself (and not just the consequences of the existing law) when it is deficient, consistent with maintaining a continuous and reasonably stable body of law; and the need to integrate law and morality (Raz 2009b).

I also believe that a sound understanding of *epieikeia* as the characteristically judicial virtue is key to resolving the problem of hard cases—cases in which the relevant existing body of law is either indeterminate or irresolvably conflicting. Such cases are correctly decided when the decision constitutes an exercise of *epieikeia*—a standard which invokes a virtue internal to legal practice, without having to posit the existence of principles implicit in the body of written law sufficient to decide every case.[14] Decisions in hard cases are subject to all the same normative requirements as decisions in cases where the law goes astray, including the requirement that the decision take the form of a novel interpretation of an existing legal text, despite the fact that the judge is actually and unavoidably making new law. Raz offers good reasons to value this interpretive exercise in this context as well. These include the need to resolve conflicts and indeterminacies in the law; and the need to give coherent shape to and make doctrinal sense of a body of written law (Raz 2009b). I am undecided whether it is too high a standard to require that judges exercise *epieikeia* in order for the authority of the State to be legitimate.[15]

7 Objections and Replies

Before proceeding to use my account of legitimate authority to resolve the problems posed at the beginning, I want to briefly address two objections to the position I have developed. Both objections concern the claim, central to my account, that the moral duty to comply with a legitimate authority's directives is ultimately grounded in the authority's interest in compliance. The first is that this claim runs counter to the spirit of the Service Conception, which my account is meant to preserve. The basic idea behind the Service Conception is that authorities are legitimate only insofar as they exercise their power for the good of those subject to their authority. If my

[14] For evidence that the ancient Athenians considered *epieikeia* to be a value internal to legal systems, see (Harris 2013, ch. 8).

[15] I hope to explore these issues more fully in future work.

account rested on the claim that what ultimately justifies authority is the benefit of exercising power to those who possess it, the account certainly would run counter to the spirit of the Service Conception. But I make no such claim. I am not arguing that an authority's directives ground a duty of compliance because he would, in his own estimation, be personally benefitted if his directive were obeyed.

The interest an authority has in having his directives obeyed is just part of his interest in maintaining social order. Because of the special social role he occupies—because he holds the position of representative of the State—an authority's interest in maintaining social order can ground a duty of compliance, while no such duty would (normally) be grounded were an ordinary citizen to start issuing commands even if those commands were conducive to social order. The authority's interest in compliance exists in virtue of the fact that he cannot succeed *qua* authority unless he obtains compliance with his directives. But the authority's interest only succeeds in grounding a duty because the authority cannot successfully serve the public unless he promotes and preserves that interest: every member of the public possesses an interest in living in an orderly society, and this is one of the interests that authorities *qua* representatives of the State are rightly entrusted to protect. An authority's interest in compliance, therefore, is simply one aspect of his interest in serving the public.

The second objection is that my account of legitimate authority commits me to the claim that an authority can only be legitimate if he views success in his work as contributing to his well-being, and that this claim is implausible. The objection proceeds as follows. I have understood a person's interests as aspects of his well-being, and have grounded duties of compliance with authoritative directives in authorities' interests in successfully and effectively serving the public. An authority who views success at his post as an important component of his own well-being clearly has the sort of interest that figures in my account. But this is not an essential feature of all authorities, or even of all authorities who satisfy the Dependence and Normal Justification Theses. And there does not seem to be any necessary connection between possessing legitimate authority and being a practical authority who views success at serving the public as an important component of his well-being.

This objection, however, relies on an assumption which I join Raz in rejecting: the Transparency Thesis (Raz 1994, p. 6). According to the Transparency Thesis, an aspect of a person's life can contribute to his well-being only if he recognizes it as doing so. Since I reject this thesis, I am not committed to the claim featured in the objection. I need not assert that authorities can only be legitimate if they view success in their work as contributing to their well-being. Rather, I need only assert that those who hold positions of authority must be said to have an interest in obtaining compliance with their authoritative acts, since failing to do so means failing in one of their central projects. And it is important to note that this recognition and respect is nowhere near sufficient to make it the case that a person of authority has been successful in his capacity as an authority. He must also exercise his authority well, in the way that is spelled out by my account. This objection, therefore, fails to make contact with my claim that duties of compliance are ultimately grounded in the interests of authorities.

Raz does claim that in order for an activity to contribute toward one's well-being, one must be engaged in that activity to some extent (Raz 1994, p. 5). I agree with this claim. This condition of engagement, however, is simply meant to exclude as contributors to well-being activities which, when performed, induce 'self-hatred, pathological self-doubt, and alienation' (Raz 1994, p. 5). If the objection aimed at my position is that it is possible for an effective *de facto* authority who satisfies the requirements of the Service Conception to suffer from one of these conditions insofar as he acts as an authority, then my response is that I find the objection fat too implausible. These conditions are, in all likelihood, incompatible with the ability to regularly and effectively serve the public interest in a position of practical authority.

8 The Problems Resolved

Having developed my account of legitimate authority, I will now proceed to argue that it has the resources to resolve the three problems for Raz's account that were posed at the outset.

8.1 Legitimate Authority and Enforceable Duties

The resolution of the first problem has already been provided in full in the course of developing the account. I have articulated the conditions which a *de facto* authority's interest in compliance must meet if that interest is to ground a justifiably enforceable duty, as well as a justifying reason to exact compliance coercively if necessary. Let us then proceed straight away to the second problem.

8.2 Theoretical and Practical Authority

The key to resolving the second problem is the claim, central to my account, that the duty of compliance owed to legitimate authorities is grounded in their interest in successfully and effectively serving the public. This is an interest that they possess *qua de facto* authorities. It is an interest which cannot be satisfied unless they obtain compliance with their authoritative directives (though this is not sufficient for its satisfaction). The interests of theoretical authorities are different. They include an interest in arriving at the truth, in assembling the most convincing and well-supported argument possible on behalf of their conclusions, and in sharing their findings. But a theoretical authority who has done all of this cannot be said to have failed *qua* theoretical authority if his findings are not translated into policy or do not end up serving as a guide for the actions of others. One has not failed *qua* climate

scientist just because one's scientific findings do not serve as an effective catalyst for the development of sensible climate change policy. This failure, rather, lies with the politicians. A given scientist may make it one of his personal goals to effectively influence policy, and count it as a personal failure if he does not do so successfully. But insofar as his central project is to do good science and not to create good public policy (as evidenced by the fact that he is working actively in his chosen scientific field and has not set aside this work in favor of taking up public office), the impact on his well-being of his failure to contribute to public policy is necessarily limited.

Since theoretical authorities do not have the same sort of interest in influencing the actions of others as practical authorities, no duty of compliance is grounded in their case. To return to our earlier hypothetical, a climate scientist who publicly ordered the American people to stop taking their cars on trips totaling less than 2 km would lack the sort of interest in compliance needed to ground a duty owed *to him* by the American people. As a matter of fact, the American people may be under such a moral duty, one which each owes to the rest of the planet's inhabitants. Such a duty, if it exists, would be grounded in the interest we all have in sustaining the earth. But this duty would have nothing to do with the interests or proclamations of climate scientists *qua* theoretical authorities.

8.3 Pre-existing Moral Duties

The problem for my account which is posed by pre-existing moral duties is that of refuting the No Difference Thesis. Raz had originally attempted to refute this thesis by arguing that the directives of legitimate authorities "turn 'oughts' into duties" (Raz 1986, p. 60, 2006, pp. 1022–1023). But this strategy will not work in cases of pre-existing duties, as opposed to pre-existing reasons that are not themselves duties. To see how my view enables me to refute the thesis, let us return to the example of Cain and Abel.

If Cain has in fact violated his agreement with Abel, then he has a moral duty to compensate him. As I have already observed, they need not be subject to any practical authority for this to be the case. I have shown in my discussion of this example, however, that the existence of such an authority does make a significant difference to their normative situation. The fact that there is an authority who satisfies the requirements of the Service Conception makes it the case that Abel would not be justified in compelling Cain to fulfill the duty owed to him on his own. Cain's duty to Abel is initially not justifiably enforceable. Once the authority has issued a ruling in Abel's favor, however, that duty becomes justifiably enforceable (by the authority on Abel's behalf). This is the normative difference which is made when an authority issues a directive in a case where a moral duty already exists. The authoritative directive changes the status of the duty, making it justifiably enforceable.

Though this should be enough to reject the No Difference Thesis, there is a further point to be made. I have argued that given the existence of an authority who satisfies the requirements of the Service Conception, someone in Abel's position

would not be justified in exacting fulfillment of the duty owed to him on his own. But in many cases, it is at least arguable that duties which would be justifiably enforced by an authority could not be justifiably enforced in the absence of one. Let us suppose that Cain has a duty to compensate Abel and that there is no authoritative person or communal body to whom they are subject. If the duty is to be fulfilled, Abel (perhaps with the assistance of his kinsman) will have to enforce it. Abel may be perfectly capable of determining how much compensation he is owed, and which of Cain's possessions would add up to that amount. But to enforce the duty owed to him justly, he must have a way of obtaining his compensation which is morally permissible and appropriate to the importance of the duty owed to him. Abel may very well have no way to do this, precisely because he lacks the resources and recognized power of an effective practical authority. It is probable that the means of collecting which are available to him will either be morally justified but unlikely to succeed, or highly effective but disproportionate and morally impermissible.[16] In cases like this, authoritative directives have an even more significant impact on the normative situations of those subject to them. They make duties justifiably enforceable which could not, practically speaking, be justifiably enforced in the absence of an authority.

My view of the relationship between pre-existing natural moral duties and legitimate authoritative directives is quite similar to Francisco Suarez's view of the relationship between natural morality and the divine law. Suarez uses "duty" to mean more or less what I mean by "natural duty," and "obligation" to mean both what I refer to as "authority-based duty" and what I refer to as "obligation" (i.e. a voluntarily self-imposed moral requirement). For Suarez, we have duties in virtue of natural facts about the world, and these duties are replaced by the obligations created by an act of God's will. If, *per impossibile* according to Suarez, God did not require and prohibit certain acts of us, we would still be duty-bound to perform, or refrain, from them; but under this scenario, God could not justly punish us for our wrongs. Suarez's understanding of the nature of God's authority, moreover, is strikingly similar to Raz's understanding of legitimate practical authority, but manages to keep the distinction between practical and theoretical authority clear. As Terrence Irwin, discussing Suarez's *Tractatus de legibus ac Deo legislatore*, explains:

> [Suarez] never suggests that obligations are the only moral requirements. On the contrary, he takes it for granted that natural facts can provide an indicative law, because they constitute reasons for us to act one way or another. Obligations in his narrow sense are not the only relations that introduce moral requirements; they introduce a different sort of moral requirement from the sorts involved in other moral relations. Obligations introduce a reason for acting that results from imposition, understood to include an expression of the will of the imposer…
>
> To impose an obligation, the imposer must be in a position to make it true, by the expression of the will for me to act in a certain way, that I have no rational alternative to

[16] We can look to history for evidence of this point. The problem of collecting on a debt, even after a court ruling, in the absence of a state mechanism of enforcement was widespread in ancient Greece. Success, which was rare, usually required a level of threats and violence which we would be reluctant to call permissible (Hunter 1993, pp. 120–150).

> acting in that way…This may be true, even if I intend to do the same action anyhow on prudential or moral grounds. The imposition of an obligation makes me aware of reasons that ought to move me even if I were unmoved by the prudential or moral grounds independent of obligation.
>
> Suarez, therefore, does not imply that without an imposed obligation we have no sufficient moral reason for observing the principles of the natural law. God's imposition gives us a further reason, but not the only reason, for observing these principles. This further reason essentially depends on God's expressing the will for us to observe these principles, not on our recognizing that God believes we ought to observe them. Hence natural law requires more than God's intellectual affirmation of the principles of natural law… [I]n [Suarez's] view, we are not obliged to do good and avoid evil before any command and prohibition. But he recognizes a moral requirement before any command and prohibition; for natural goodness and badness tell us what we ought (*debere*) to do. Divine commands, introducing genuine law, oblige us to do something that we already ought to do. Natural law, therefore, requires both a divine command and prior intrinsic rightness. The moral judgment and the recognition of moral duty (*debitum*) are prior to any act of will, by the lawgiver or by the subject. In all this discussion Suarez distinguishes 'ought' (*debere*) and 'duty' (*debitum*) from 'obligation' (*obligatio*). A class of actions that we ought to do, and that it would be right to do and wrong to avoid, is already fixed by nature; the divine command adds an obligation to do the things we already ought to do. (Irwin 2008, pp. 21–22, 29–30)

9 Conclusion

We now have at our disposal a detailed account of the moral grounds of individual rights and duties, and of the powers of legitimate authorities over the individuals in their jurisdictions. We are thus properly situated to address the question of whether the State does indeed have the authority to create and maintain the conditions of Equal Liberty, and if so, what makes this the case. This question has a counterpart: the question of what principle, if any, limits the authority of the State to intervene in the lives of individuals, and what grounds such a principle. The answer I will offer, in brief, is that when correctly interpreted, the Harm Principle—which is the proper limit on State authority—leaves the State exactly the space to act that it requires in order to achieve the goals of Equal Liberty. I will therefore postpone the application of the theory of authority just developed, until after we have examined some important, but ultimately unsuccessful, attempts to delineate the appropriate realm of State intervention. I will then address together the tasks of laying out the ground of the State's authority to pursue its distributive aim and the ground of the proper limitation on State action.

References

Aristotle. *Nicomachean ethics* [Ethica Nicomachea]. Oxford classical texts, ed. L. Bywater. Oxford: Oxford University Press.

Aristotle. *Politics* [Politica]. Oxford classical texts, ed. W.D. Ross. Oxford: Oxford University Press.
Dancy, J. 1993. *Moral reasons*, 127–143. Oxford: Blackwell.
Darwall, S. 2009. Authority and second-personal reasons for acting. In *Reasons for action*, ed. D. Sobel and S. Wall, 134–154. Cambridge: Cambridge University Press.
Edmundson, W.A. 1998. *Three anarchical fallacies*. Cambridge: Cambridge University Press.
Harris, E.M. 2013. *The rule of law in action in democratic Athens*. Oxford: Oxford University Press.
Hunter, V.J. 1993. *Policing Athens: Social control in the Attic Lawsuits, 420–320 B.C*. Princeton: Princeton University Press.
Intergovernmental Panel on Climate Change. 2007. *Climate change 2007: The physical science basis*. Cambridge: Cambridge University Press.
Irwin, T.H. 2008. *The development of ethics: A historical and critical study. Volume II: From Suarez to Rousseau*. Oxford: Oxford University Press.
Miller, F.D. 1995. *Nature, justice, and rights in Aristotle's politics*. Oxford: Clarendon Press.
Quong, J. 2011. *Liberalism without perfection*. Oxford: Clarendon Press.
Raz, J. 1975. Permissions and supererogation. *American Philosophical Quarterly* 12(1): 161–168.
Raz, J. 1979. *The authority of law*. Oxford: Oxford University Press.
Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.
Raz, J. 1994. Duties of well-being. In *Ethics in the public domain: Essays in the morality of law and politics*, 3–28. Oxford: Clarendon Press.
Raz, J. 2006. The problem of authority: Revisiting the service conception. *Minnesota Law Review* 90: 1003–1044.
Raz, J. 2009a. Intention in interpretation. In *Between authority and interpretation: On the theory of law and practical reason*, 265–298. Oxford: Oxford University Press.
Raz, J. 2009b. Interpretation: Pluralism and innovation. In *Between authority and interpretation: On the theory of law and practical reason*, 299–322. Oxford: Oxford University Press.
Sherman, J. 2010. Unresolved problems in the service conception of authority. *Oxford Journal of Legal Studies* 30(3): 419–440.
Simmons, A.J. 1979. *Moral principles and political obligations*. Princeton: Princeton University Press.

Chapter 16
The Scope and Limits of State Authority

1 Introduction

A liberal society must be no less interested in setting appropriate limitations on State power than it is in enabling public action for the sake of achieving distributive justice. The other side to our investigation of the proper moral grounds of political authority, therefore, is an investigation of these limitations. John Stuart Mill presents the Harm Principle as a strict constraint on State interference in the lives and actions of individuals (Mill 1869/1978). The utilitarian argument with which he supports the principle, however, gives to thelife while still legitimately term "harm" an alarmingly wide range of meanings. The argument leaves open the possibility of advocating extensive state control over individual life while still legitimately claiming to endorse Mill's liberal political philosophy. This possibility runs counter to the spirit of liberalism, and excluding it is one of the goals of Joseph Raz's alternative autonomy-based argument for the Harm Principle (Raz 1988). But Raz's version of moral perfectionism prevents him from seeing autonomy as intrinsically valuable, and his argument remains vulnerable to a commitment to extensive State control, if only this control can be achieved in a sufficiently efficient and subtle way.

I begin the chapter by discussing Mill's utility-based argument and some of its most severe defects. I follow this with a discussion of Raz's theory, and an extended argument against some popular reasons for believing that Razian perfectionism is illiberal. I then elicit the real problem in Raz's theory, which I call "the efficiency problem." Raz makes a few brief indications as to how he might respond to this problem, but none of these is sufficient to diffuse it. I argue that given Raz's understanding of the nature and value of autonomy, he cannot avoid the susceptibility to extensive State control that he identifies in Mill. I then advance my own liberty-based interpretation and defense of the Harm Principle, using the conception of liberty developed in this book. I begin by considering the nature of the value of liberty. I then provide an account of the ground of the State's authority to establish the conditions of Equal Liberty. Finally, I argue that a liberty-based version of the

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0_16

Harm Principle is the appropriate limit on the exercise of this authority, and that my defense of this principle is not vulnerable to the same objections as Raz's.

2 Mill's Utility-Based Argument and Its Defects

Mill claims that the "sphere of action…comprehending all that portion of a person's life and conduct which affects only himself, or, if it also affects others, only with their free, voluntary, and undeceived consent…is the appropriate region of human liberty" (Mill 1869/1978, p. 71). This region is divided into three parts: one's own thoughts and the expression of them, one's tastes and plans for the development of one's life and character, and one's associations with other people (Mill 1869/1978, p. 71). Within this region of human action, interference from the State should be limited to the prevention of harm to other people. Mill argues this point by first establishing that these liberties generally tend to increase total utility, and then arguing that they only generally tend to decrease utility when they result in harm to others. Rather than attempt to define harm at the outset, we should allow Mill's understanding of harm to emerge from the development of his argument by identifying the kinds of actions the argument recognizes as legitimate targets of State interference.

Mill does not give special attention to the freedom of association, since this freedom is in the service of the other two. If people wish to come together in a group, their purpose is either to express and exchange their thoughts on some matter, or to pursue some activity in accordance with their tastes or plans, or both of these. One could hardly claim to have defended the liberties of expression and planning one's life if the defense is limited to expression in the absence of other people, and to plans that can be achieved in total isolation. An adequate defense of the first two liberties will thus accomplish a defense of the third.

As a utilitarian, Mill's argument for the Harm Principle must be grounded in the principle of utility, and he asserts early in his essay that this latter principle is "the ultimate appeal on all ethical questions; but it must be utility in the largest sense, grounded in the permanent interests of man as a progressive being" (Mill 1869/1978, p. 70). His argument for freedom of expression is based on epistemic utility. Knowledge is both intrinsically and instrumentally useful: the acquisition of knowledge is itself one of the greatest pleasures humans can experience, and the application of knowledge can result in great material benefit for a great number of people. A suppressed idea may be wholly true, partially true, or false. If it is true in whole or in part, the disutility of its suppression is obvious: we can neither enjoy its possession nor benefit from its application. Mill considers the objection that State officials must assume their views are right in order for those views to guide their conduct, and once they have made this assumption they have sufficient reason to suppress opposing views that may lead to subversive conduct. For rational people, however, this assumption of rightness must be justified, and this justification is only attained when there is "complete liberty of contradicting and disproving our

opinion" (Mill 1869/1978, p. 79). Only a view that can be defended in such a hostile environment deserves to guide our actions.

The epistemic utility of asserting some false propositions is also very great. Mill asserts that "the cultivation of the understanding…is surely in learning the grounds of one's own opinions" (Mill 1869/1978, p. 97). Seeking to understand the justification of the truths one is taught fosters "that generally high scale of mental activity which has made some periods of history so remarkable" (Mill 1869/1978, p. 96). To accept what is taught without seeking its justification produces mental laziness on the part of the student. This laziness decreases the likelihood of discovering new truths, since discovery depends so heavily on intellectual vigor. We cannot fully understand the justification for a view, or have complete confidence in its truth, until we have identified the flaws in all the objections to that view. The objections must therefore be publicly stated and argued as vigorously as possible. Nor can we ever declare that some matter has been settled once and for all.[1] There is no way to determine whether every objection has been answered, since new ones may always be contrived in the future.

Freedom of expression is therefore generally beneficial to the well-being of societies and the individuals that constitute them. The principle of utility will only allow this freedom to be infringed in cases in which this general benefit is outweighed by some disutility. For Mill, this can only happen when the disutility is a harm to another person. Again, I am postponing the discussion of what exactly this means until the end of the examination of Mill's argument. Let us proceed then to his argument for freedom in the planning of one's own life.

Mill's key observation in this part of the argument is that human beings differ greatly in their "sources of pleasure, their susceptibilities of pain, and the operation on them of different physical and moral agencies" (Mill 1869/1978, p. 133). These differences necessitate the availability of a wide range of life-plans. Those who are not able to live the kind of life that suits the idiosyncrasies of their natures will not "obtain their fair share of happiness, nor grow up to the mental, moral, and aesthetic stature of which their nature is capable" (Mill 1869/1978, p. 133). In general, we maximize utility—both our own and what we contribute to the lives of others—when we live lives that suit our individual natures. The principle of utility at least demands the promotion of individuality. The freedom to plan one's life, however, is more than the freedom to live the life one is most suited to. It is the freedom to choose for oneself what kind of life to lead, even at the risk of choosing a life that does not develop one's nature and fulfill one's potential. Mill gives three arguments for why this freedom tends to maximize utility: an epistemic argument, an argument from personal dedication, and an argument from the significance of choice.

[1] Mill claims that the truths of geometry are an exception, since there is nothing intelligible to be said against them. Why he thinks this is not entirely clear, and had he been alive during the twentieth century, he would have appreciated the possibility of equally skilled mathematicians disagreeing on the truth-value of certain mathematical propositions (e.g. any proposition whose proof requires use of the axiom of choice).

The epistemic argument has two parts. First, in many cases the individual is by far the best judge of what kind of life will be most suited to his character. Mill acknowledges the importance of education and the value of humanity's past experiences in helping each of us determine what kind of life is worthwhile. Only the individual himself, however, can "find out what part of recorded experience is properly applicable to his own circumstances and character" (Mill 1869/1978, p. 122). No one can acquire greater familiarity with one's own character and circumstances than oneself. One is thus naturally in the best epistemic position to determine, from a range of worthwhile options, what kind of life will yield the greatest happiness for oneself and other people.

Nonetheless, there are some people who are so bad at determining what kind of life would best suit them that they reject all worthwhile options and choose a life that is of no good or use to anyone, themselves included. Why should these people not be forced to pursue some worthy project? Mill claims that when the public interferes in purely personal behavior "odds are that it interferes wrongly and in the wrong place" (Mill 1869/1978, p. 151). The State or the public in general would only be justified in interfering if it could be counted on to improve the situation. We should not be so quick, however, to accept this reply. It seems unlikely that there is no class of cases in which the State is competent to judge that those whose lives belong to that class contribute nothing positive to society. It does not seem a mere prejudice to say that someone who devotes all his available free time to the consumption of alcohol or recreational drugs, working just enough to supply himself with his chosen diversion, is an unproductive member of society. Mill must argue that the principle of utility supports the right to *waste* one's life, provided that doing so does not result in harm to others. Since there seem to be some cases in which not much doubt attaches to whether a life is being wasted, this conclusion will have to be supported by his other two arguments.

The argument from dedication to one's projects gets us some way toward the desired conclusion. Mill observes that "A person whose desires and impulses are his own…is said to have character…If, in addition to being his own, his impulses are strong and under the government of a strong will, he has an energetic character" (Mill 1869/1978, p. 124). The possession of an energetic character is important from the viewpoint of utility because "more good may always be made of an energetic nature than of an indolent and impassive one" (Mill 1869/1978, p. 124). If the State were to coerce those who waste their lives into pursuing valuable projects, it is unlikely that they would make any significant progress toward accomplishing them. They would be alienated from their impulses to pursue these projects, since they would be responding to the coercion rather than the value of what they are working toward, and their impulses would be weak to begin with. The utility of forcing them to make some small contribution would have to outweigh the cost of implementing the coercion.

We can make two responses to this argument. The first is to deny the accuracy of Mill's initial observation. It probably is true that many people will act from an "indolent and impassive nature" when they are first coerced into doing something. It is possible, however, that if the project really is valuable they will come to embrace

it over time (this possibility will play an important role is Raz's argument). We should not assume that the implementation costs will never be outweighed by the benefit. The second is to say that even if most people cannot be coerced into making a significant contribution with their lives, perhaps some can. There may be some for whom this kind of impetus is all they need to become productive members of society. As a rule utilitarian, it might be enough for Mill if most people do not fit into this class. The best policy would then be one of non-coercion, provided that the implementation costs of coercion are significant. If the State could determine, however, how likely someone was to embrace eventually a project they were forced into, rule-utilitarianism would have to allow coercion in that class of cases, unless there were an additional countervailing reason. Providing such a reason is the purpose of Mill's third argument.

Mill's argument from the significance of choice attempts to respond to the suggestion that someone might eventually embrace a way of life into which he was coerced. Mill tries to exclude this possibility by insisting on a connection between energy and choice. He claims that we exercise our faculties only when we make free choices (Mill 1869/1978, p. 122). A project can only be pursued energetically if it engages our faculties, because "if the inducements to an act are not such as are consentaneous to his own feelings and character…it is so much done towards rendering his feelings and character inert and torpid instead of active and energetic" (Mill 1869/1978, p. 123). Thus, we can only put our energy into plans we have chosen. When one is coerced, one fails to deploy one's faculties "other than the ape-like one of imitation" in the pursuit of one's project (Mill 1869/1978, p. 123). An activity that is not of one's choosing will necessarily fail to engage one's nature. For Mill, this initial lack of engagement rules out the possibility of the project being embraced later on. The development of an individual human nature must take place "according to the tendency of the *inward forces* which make it a living thing [emphasis added]" (Mill 1869/1978, p. 123). To develop as people, the projects we pursue must be set by our own internal natures. In asserting that "in proportion to the development of his individuality, each person becomes more valuable to himself, and is, therefore, capable of being more valuable to others" he indicates his belief that one cannot come to value a project into which one is coerced (Mill 1869/1978, p. 127). Since a coerced project is not part of the self-development of one's individuality, one cannot value the work one contributes to it. If the project is not valued, we cannot expect it to be pursued energetically. This makes it very likely that the cost of implementation will outweigh the increase in utility gained from the pursuit of the coerced project.

We need not dwell long on the flaws of this argument. Mill must assume that the performance of the activity itself, even if initially resisted, could not eventually spark the interest of the one coerced, and through this interest lead to a willful acceptance of the project. There is no reason to assume this kind of obstinacy in human psychology. Mill fails to block the State's justification for interfering in purely personal behavior in all cases. The principle of utility not only justifies state interference in some (albeit specific and limited) kinds of personal behavior, it allows for a very broad definition of what constitutes a harm to another person. We

know that the liberties of expression and of planning one's life are generally supported by the principle of utility. Since Mill is a rule-utilitarian, the Harm Principle cannot allow for State interference in just any instance in which allowing someone one of these liberties will result in a decrease in total utility. Harms must be that class of acts that generally tend to decrease utility, to lead to less total happiness or pleasure.[2] The decrease in utility must, moreover, be felt by people other than the agent. Mill provides two definitions of this class of actions. He first tells us that a harm is an action that "violate[s] a distinct and assignable obligation to any other person or persons" (Mill 1869/1978, p. 148). An obligation of this special sort has been violated whenever an action leads to "a definite damage, or a definite risk of damage, either to an individual or to the public" (Mill 1869/1978, p. 149). The term "definite damage" requires explication, which we are now in a position to provide. Whenever someone acts, the consequences of the action fall into one of five possible categories: utility-maximizing, imperfect utility-increasing, neutral, imperfect utility-decreasing, or utility-minimizing. Given Mill's attempt to defend one's right to waste one's life, the Millian definition of a "definite damage" must be a consequence that falls into either the fourth or the fifth category (and usually the fourth, given the rarity of someone's doing an action which, given the circumstances, causes the most pain that could possibly be caused). Actions that fail to produce any utility, or produce very little, do not count as harm, even when these actions fail to prevent easily preventable disutility to other people.

We have seen that Mill fails to establish the claim that the principle of utility forbids State interference in the lives of those whose actions are mainly neutral or productive of very little utility. His definition of harm itself yields a wide range of acceptable cases of official interference. To say that no assignable obligation has been violated, or no definite damage done, is to say that the action of an individual "neither violates any specific duty to the public nor occasions perceptible hurt to any assignable individual except himself," nor does it pose any clear risk of having one of these consequences (Mill 1869/1978, p. 149). We must add this point about risk, since Mill has already identified the risk of definite damage as a condition that legitimates State interference. But there are many actions that pose a risk of causing perceptible hurt to one or more individuals. Mill cites gambling, drunkenness, idleness and uncleanliness as acts that fall *outside* the realm of what should be prohibited by law (Mill 1869/1978, p. 147). But drunkenness brings with it the possibility of injury to other people, and gambling by someone with a family to support constitutes a violation of an assignable obligation. Severe uncleanliness in public can contribute to the spread of disease. Even idleness can cause perceptible harm to the public, if the State has any kind of welfare system. Mill makes explicit that he views certain causes of taxation and public expenditure as perceptible harms to the public. He asserts that if one has a child, one should be compelled by the State, as far as is possible, to feed, clothe and educate it at one's own expense. Anything less is a

[2] Though Mill, unlike Bentham, allows for the existence of higher and lower pleasures, he still endorses the Benthamite equations of utility = happiness = pleasure and disutility = unhappiness = pain.

crime against society as well as the child (Mill 1869/1978, p. 176). Mill's argument certainly does not allow us to view all taxation as harm. Some services necessary to the well-being of every individual in a society, such as police and fire departments, can only be provided by the State. Some welfare programs are also justifiable. Whenever someone requires financial assistance as a result of circumstances beyond his foresight or control, the principle of utility demands that the public as a whole bear the cost of helping him. But idleness and gambling by those who cannot afford them, and private drunkenness that results only in injury to oneself, provided that the injury must then be treated by publicly funded medical care, are actions that, at least initially, lie within our control. They are the kinds of actions, moreover, that generally tend to result in financial demands on the public, at least in a society with the kind of welfare system that the principle of utility would demand. The only alternative to providing the idler, the drunk, and the gambler with some basic care and support is for the state to turn a blind eye and simply let them suffer the consequences of their actions. This is unjustifiable from a utilitarian point of view, given that the amount of suffering that would result can hardly be expected to be outweighed by the inconvenience of a slightly higher tax burden spread over a large population.

By allowing the State to interfere based on the risk of harm, and counting the burden of taxation as a harm to the public when it results from actions that could have been avoided, Mill's theory shrinks the freedom to plan one's life, and implies the acceptability of coercion even in some cases that he explicitly wants his theory to place outside the realm of State authority. One might object to my extension of Mill's remarks on the duties of parents into the cases of the idler, drunk, and gambler. The parent has a duty to another person, and if he or she fails at it, the State has the right to intervene. But in the other cases, although there is the risk of perceptible harm to others, only duties to self are being violated. This objection misses the point. For Mill, the duties we have to ourselves are the duty to preserve our dignity and the duty to develop our natural abilities (Mill 1869/1978, p. 145). He does claim that the Harm Principle prevents the State from forcing anyone to perform these duties. I have not, however, been arguing that the Harm Principle allows for the coercion of the idler, private drunk, and gambler on account of their lack of dignity or unfulfilled potential. If we suppose that they are independently wealthy, and thus in no need of societal support, there is no way to advocate coercing them based on the risk they pose to others. But once they become the source of financial harm to the public, the fact that this harm results from a violation of purely personal duties is irrelevant. For Mill's principle, only the resulting public harm matters. As soon as the risk of "perceptible hurt" is present, interference is justified.

Mill's broad definition of harm places severe restrictions on the freedom to plan one's life as one chooses, including some restrictions he explicitly denies endorsing. Because Raz's refashioning of Mill's argument aims primarily at resolving Mill's problems in establishing this freedom, I will say relatively little about the impact of the definition of harm on the freedom of expression. Mill certainly gets the desired result of being able to restrict expression that poses a risk of causing immediate harm to the health or interests of other people. His broad understanding of harm,

however, is evident in this region of freedom as well. He allows for the prohibition of "violation[s] of good manners" and "offenses against decency" (Mill 1869/1978, p. 168). This may place some limitation on the vigor with which ideas are allowed to be debated.

3 Raz's Autonomy-Based Argument

Joseph Raz refashions the argument for the Harm Principle in a way that seeks to avoid the difficulties inherent in Mill's defense of the freedom to plan one's own life. Rather than argue from the principle of utility, Raz makes an argument from the reality of competitive moral pluralism and the value of autonomy. After examining his development of this argument, I will comment on the ways in which it improves on Mill's, and then discuss the problems that remain for it.

According to Raz, to be autonomous is to be a part author of one's life. This requires that "many morally acceptable, though incompatible, forms of life be available to a person" (Raz 1988, p. 158). The available forms of life must be incompatible, because being a part author of one's life requires having a range of choices concerning the kind of life one will lead. Without incompatible options, there is no need to make a choice about one's life, and one cannot be an author of one's life if one never has an opportunity to choose one way of life over another. Autonomy is something more than the lack of coercion. It is possible for someone to have only one form of life available to him and, as luck would have it, to be enthusiastic about leading that kind of life. Such a person has not been coerced, but he is not autonomous. Everything has been set for him whether he likes it or not. Raz stresses that autonomy requires an *adequate* range of options, but does not require any one option. The key is that of all the forms of life available to an individual, some group of them be lives he actually wants to lead. But one can be autonomous even if the way of life one wants most of all is not available.

The range of available forms of life must be morally acceptable as well as adequately diverse. A range of options that does not include any morally acceptable ones is obviously inadequate, since many people will fail to find in such a range any life that they want to lead. The real point of this moral requirement is in what it excludes. The presence of morally repugnant options is not a requirement of autonomy (Raz 1988, p. 169). Raz acknowledges that morally good individuals are "able to cope with the temptations and pressures normal in their society," and thus that "vices and moral weaknesses are logically inseparable from the conditions of a human life that can have any moral merit" (Raz 1988, p. 168). He does not, however, think it is possible to eliminate these problems, and so it is no requirement of respect for autonomy to ensure that they are available. We shall return to this point when we examine the difficulties faced by Raz's argument.

Raz does not believe that autonomy is valuable in itself. An autonomous pursuit of evil or vice is more worthy of blame than a similarly evil but coerced pursuit. The value of autonomy lies in choosing to pursue a good form of life (Raz 1988, p. 169).

A good life, according to Raz, is a life of self-realization, a life that develops and exercises some of one's valuable capacities. Autonomy is not required for self-realization, and being autonomous means being able to choose a life of self-realization or reject it. Autonomy itself has no value if the wrong choice is made. An autonomous life of self-realization, however, is better than a non-autonomous self-realizing life. In accordance with the condition of autonomy that incompatible forms of life be available, there must be many different, incompatible ways of self-realizing. We cannot develop all our capacities fully, so it must be morally acceptable and consistent with autonomy to have to forgo developing some in favor of others. This requirement forges a connection between autonomy and value pluralism. A morally acceptable form of life exhibits certain virtues to a high degree and others to a lesser degree. For there to be incompatible morally acceptable forms of life, there must be incompatible virtues. Choosing to lead a life that exhibits one set of virtues to a high degree means giving up a life that exhibits others virtues. The different virtues exhibited by these forms of life provide the competing reasons for choosing them. Autonomy requires that available forms of life be *equally* morally acceptable. Being autonomous requires being able to make one of several choices, none of which is worse overall than the others. Otherwise, there would be a sense in which the situation made the choice for one, provided one was committed to leading the most morally valuable life one could.

The link between respect for autonomy and toleration is made through competitive value pluralism, a view to which both Raz and I are committed:

The Principle of Competitive Value Pluralism: There are many equally good ways of life, which are incompatible insofar as leading one excludes leading others, and the values that structure some conflict with the values that structure others.

This view "admits the validity not only of distinct and incompatible moral virtues, but also virtues which tend, given human nature, to encourage intolerance of other virtues" (Raz 1988, p. 164). Raz claims that in addition to tolerating behavior we view as bad, we can also tolerate people's limitations, when these limitations result from the choice to develop certain virtues at the expense of others. If one has chosen a life that allows one to cultivate the virtues of decisiveness and expedience in action, one will probably be tempted toward intolerance of those whose lives have cultivated the virtues of cooperation and careful deliberation. One's awareness of the different but equal value of this other form of life, which is incompatible with one's own, gives one a reason to tolerate the other person's limitations, and to expect that one's own limitations will be tolerated.

If we respect autonomy as one component of a good life, we must value moral pluralism. The diversity of human virtues makes it likely that this pluralism will be competitive. Our respect for autonomy thus provides a reason to value toleration. The final step in Raz's argument is to show that the appropriate principle for limiting the scope of autonomy-based toleration is the Harm Principle. Raz distinguishes his version of the Harm Principle from Mill's defining it as one which "regards the prevention of harm to anyone ([the agent] included) as the only justifiable ground for interference with a person" (Raz 1988, p. 169). Raz gives a narrow and a broad

sense to the notion of harm. In the narrow sense, someone is harmed when his prospects are limited or his efforts toward a project he has already begun are frustrated (Raz 1988, p. 169). In the broad sense, harm includes physical injury that is not incapacitating and reasonably endurable offense (Raz 1988, p. 170).

The argument from autonomy is an argument for the narrow understanding of the Harm Principle. Harm in the narrow sense just is restriction on autonomy. This is one great improvement over Mill's argument from utility. Many of the acts that must count as harms to the public based on Mill's argument will not so count on Raz's, because they do not meet the more stringent requirement of restricting the autonomy of others. The argument from autonomy thus preserves more of the spirit of liberalism that the Harm Principle is supposed to embody. Raz's argument also provides justification for a broader system of welfare than Mill allowed. Raz identifies three duties essential to the promotion of autonomy (Raz 1988, p. 166). The first is the negative duty of refraining from coercing others unless coercion is needed to prevent a harm. The other two are positive duties: we must help others cultivate the capacities that a good life requires, and help make an adequate range of good lives available to others. Basing the Harm Principle in a respect for autonomy means "establish[ing] that the autonomy-based duties never justify coercion where there was no harm" (Raz 1988, p. 171).

Raz considers the objection that the existence of positive autonomy-based duties requires the State to coerce its citizens into contributing to the autonomy of others (most likely through compulsory taxation), and that the failure of someone to make such a contribution does not constitute a harm. His reply is that this objection relies on an unjustifiably narrow definition of harm, and he asserts that "one can harm another by denying him what is due to him" (Raz 1988, p. 171). Because autonomy is valuable when it is part of a life of self-realization, we have a duty to promote it by making acceptable options available to others. If we fail in this, other people will have fewer prospects than if we had succeeded—fewer prospects than they *should* have had, since the value of autonomy makes its promotion a *duty*. The central meaning of harm is limiting someone's prospects. We harm others when we fail to fulfill our autonomy-based duties, because our actions result in some people having an insufficient range of valuable prospects, and thus in those people being denied the possibility of good lives. The existence of State-enforced autonomy-based duties, therefore, does not contradict the Harm Principle.

4 Failed Objections to Razian Perfectionism

There are some serious problems with Raz's autonomy-based version of the Harm Principle, including the fact that he does not have an adequate account of the ground of these duties of autonomy, or of the State's claim-right against its citizens that they perform these duties. But before we address the very real problems in Raz's theory, let us examine some important objections to it that do not ultimately succeed.

Jonathan Quong has recently put forth a spirited defense of Rawlsian, anti-perfectionist liberalism (Quong 2011). He makes four arguments against Razian perfectionist liberalism. The first is essentially that Raz's theory of legitimate practical authority fails to explain why such authorities have a claim-right to compliance with their directives. We have already seen that this is a sound and important objection to Raz; but Quong goes wrong in assuming that it is a problem that no Razian theory of authority can solve. As we saw in the last chapter, once Raz's theory of authority is embedded within my broader theory of moral rights and duties, the problem can be readily solved. So Quong's first argument need not detain us.

The other three arguments may be termed "the manipulation argument," "the paternalism argument," and "the contingency argument." All three of these arguments have the same goal: to show that Razian perfectionism does in fact violate the Harm Principle, and is thus illiberal. In this section, I examine the first two of these arguments, and show that neither stands up to scrutiny. In the following section, I discuss the contingency argument, which introduces a problem in Raz's theory quite close to the one I have elsewhere referred to as the efficiency problem (Sherman 2007). This is a serious objection to Razian perfectionism, and one which I doubt Raz's own theory has the resources to solve. As we will see in the second half of this chapter, however, my version of liberal perfectionism, by providing the basis for a liberty-based interpretation of the Harm Principle and forcing us to rethink the nature of the value of liberty, does.

4.1 The Manipulation Argument

The manipulation argument relies on Robert Nozick's explanation for the fact that to threaten someone is an act of coercion and thus a violation of autonomy, while to make someone an offer is not. In brief, Nozick's position is that a threat places someone in a choice-situation which he would not (if he is rational) have chosen to place himself, and thus whatever choice he makes after the threat has been made—the choice to comply or the choice to refuse—is not autonomous. An offer, on the other hand, places someone in a choice-situation which he would, rationally, choose to place himself, since it provides a choice to either accept or to reject and thus maintain the *status quo*. So offers do not involve any violation of autonomy.[3]

Quong's anti-perfectionist argument concerns the sorts of tax-funded subsidies which a perfectionist State uses to expand access to valuable activities—a type of policy which we may assume, in all that follows, violates either intention-neutrality or argument-neutrality (or both), as the value of these activities and the cogency of the arguments for them are not commonly recognized. He does *not* argue that the

[3] This latter claim can be seen to be quite obviously wrong, when we consider cases of offers made by someone who occupies a position of power over the one to whom the offer is made, to engage in some activity desired by the offerer but not by the offeree. This fact appears to escape both Nozick and Quong, but it is not essential to my present argument.

perfectionist State makes any threats against its citizens. However, he claims that taxation for the purpose of such subsidies is a form of manipulation, and that this too is a violation of autonomy. Manipulation is a non-threatening way of placing individuals in choice situations which they would not rationally choose to place themselves. Let us examine the key passage in Quong's argument:

> The choice for citizens is between having the money to spend themselves, or having the government take it from them and then spend it on subsidizing [for example] opera tickets. Since the latter option simply reduces what you can do with your resources, it would be irrational to prefer it, and thus we cannot construe perfectionist subsidies as unproblematic offers in Nozick's sense. We must assume (subject to a caveat discussed below [i.e. the case of genuine public goods, like mass transit]) that citizens would prefer the status quo (keep their resources) over the post-subsidy situation where the government taxes them and uses those funds to subsidize the opera. By putting citizens in the post-subsidy situation, the government thus does attempt to subject the will of citizens to its own perfectionist judgement. Perfectionist subsidies are not like the offers one person might make to another; they involve the government taking funds from citizens in order to restrict the ways in which citizens can spend those resources. Nozick's distinction thus supports the view that perfectionist subsidies are a form of autonomy intrusion since, under normal conditions, they represent the government placing citizens in a choice situation in which they would not have chosen to place themselves, and so citizens' subsequent choices would be, on Nozick's account, not fully their own: they represent an attempt to subject citizens to the will of the perfectionist state. (Quong 2011, pp. 65–66)

There are two problems with this argument. The first is that there is absolutely no justification for the claim, which Quong fails to defend or even to question, that it would be irrational to prefer the option of paying taxes to fund subsidies which broaden access to valuable activities within one's community, since "the latter option simply reduces what you can do with your resources." There is nothing irrational about such a preference whatsoever. For an individual to rationally hold such a preference, he need only prefer to live in a society in which there is broadly distributed access to the fine arts—a preference which he may rationally have, for any number of reasons—and recognize that the most efficient way of realizing that goal is through government action which he has the power to support *qua* tax-payer—an eminently rational belief.

If Quong's argument is to survive this obvious difficulty, we must interpret him as claiming that this sort of government action is manipulative assuming that individuals do not wish to subsidize broader access to the arts, whether or not that is rational on their part. We might question whether this alteration leaves Quong with a defensible concept of manipulation at all, but that is not the point I wish to pursue here. Let us assume that the mere fact that (some) individuals do not wish to fund these subsidies makes them *prima facie* manipulative. The question of whether they are actually manipulative then rests on an issue which, in his discussion of the ways a Razian might respond to his argument, he fails to consider. Taxation to fund such subsidies cannot count as manipulation if it is the most efficient and least intrusive way for the State to fulfill a duty which it has to its citizen body, and if the State has a right that all citizens cooperate with its efforts to fulfill this duty. In such a scenario, the State is simply choosing the best available means to getting its citizens to

do what they have an enforceable duty to do. And getting someone to do what he has an enforceable duty to do *cannot* count as manipulating him. I think it is perfectly plausible that access to the arts must count as part of what is required for an individual to possess an acceptable amount of autonomy-freedom in a developed society. As such, it falls within the scope of the policy goals of Equal Liberty. In the second half of this chapter, I will demonstrate that cooperation with its pursuit of the policy goals of Equal Liberty is *precisely* what the State has a right to demand from its citizens. We will thus be forced to conclude that Quong's manipulation argument fails.

4.2 The Paternalism Argument

Let us turn, then, to the paternalism argument. Quong's first claim concerns one way of distinguishing perfectionist and anti-perfectionist liberal theories:

> What…perfectionism must claim, in order to practically distinguish itself from theories such as Rawls', is that even if everyone has been given their fair share of rights, liberties, opportunities, income, and wealth, further perfectionist policies will be necessary. Although this claim is an essential part of all contemporary theories of perfectionism, it is seldom explained. Why should state action be required even if resources have been fairly distributed to individuals? (Quong 2011, p. 85)

The aim of the paternalism argument is that any response a perfectionist liberal can give to this question will commit him to a morally objectionable form of paternalism. Quong defines paternalism as follows: "[T]he core element of paternalism [is] the paternalizer holding a negative judgement about the paternalizee's capacity to effectively advance his or her own interests" (Quong 2011, p. 83). One response Quong considers concerns weakness of will (Quong 2011, p. 88). Perhaps many of the individuals in a given society judge that they should engage in valuable activities of a certain sort, but fail to do so owing to weakness of will. Quong claims it is paternalistic to enact policies designed, on the basis of this assumption, to make it more likely that these individuals will engage in the valuable activities by altering their choice situations to make it easier for them to overcome this weakness. Such action by the State, he insists, implies that the sort of negative judgment characteristic of paternalism has been made.

The only suitable response to this claim is to point out that we know, empirically, that a diminished degree of control over one's decisions due to ego-depletion is a real psychological phenomenon.[4] There is a strand of liberal thought, in which I think Quong must be placed, that wants to pretend that this is not so—that wants social policy to be guided by the demonstrably false assumption that the unmanipulated actions of a mature and healthy adult are always expressive of his judgments of what is in his own best interest. Such thinkers do not necessarily believe that this is true; but they do believe that we cannot avoid morally objectionable

[4] See the discussion of self-control and weakness of will in Chap. 7.

violations of autonomy unless we act as if it is. But this belief is groundless. If we can understand the structures of the types of choice-situations which, as a matter of fact, are especially conducive to ego-depletion, and enact policies which prevent these situations from arising, and replace them with choice-situations structured to be favorable to exercising self-control, we will have done something which leaves individual *more* free to act autonomously, not less. Self-control, after all, is one of the dimensions of autonomy. This is precisely the sort of social research and policy program which Richard Thaler and Cass Sunstein have recently advocated (Thaler and Sunstein 2008). The major flaw in their work is one of labeling. They describe what they are advocating as "soft paternalism."[5] But what they advocate is not a form of paternalism at all. Given an appropriate understanding of individual autonomy, these sorts of policies are autonomy-enhancing.

Policies which make us freer to act autonomously *cannot* be paternalistic, if the term "paternalism," is to have anything like its traditional and commonly understood meaning, and if it is to pick out a form of behavior which is at least presumptively morally objectionable. Quong's characterization of paternalism, then, must be incomplete. To count as a paternalist, one cannot merely motivated by a negative judgment of the kind Quong describes. Such judgments are often correct; and it simply *cannot* be the case that in order to act ethically, we must replace accurate judgments about the world with purported belief in demonstrably false views of human nature. The paternalist is one who is motivated by this sort of judgment *to act so as to further hinder the individual's ability to act autonomously*, by simply replacing the individual's judgment with his own and either coercing or manipulating the individual into doing what he thinks is best. Perfectionist policies motivated by a recognition of the fact of weakness of will need not do this. They may do quite the opposite, by creating a choice-environment in which agents are more likely to succeed in acting autonomously.

Quong considers three other responses on behalf of the perfectionist: that perfectionist policies are required because (at least some) individuals are (at least sometimes) not rational, in the sense of not being able to determine what is in their best interest; that they are required because without them, the cost of many valuable activities would be prohibitive for many individuals; and that they are required because having certain sorts of experiences is required for leading a good life, and these policies make it more likely that more people will have these experiences. I group these together because Quong's objections to these imagined responses are all based on the same mistakes. I will briefly sketch the individual objections, and then identify the errors on which they rely.

Quong's objection to the rationality response is essentially the same as his objection to the weakness of will response: State action motivated by the judgment that an individual is not able to determine what is in his own best interest amounts to a morally objectionable paternalism (Quong 2011, p. 86). His objection to the prohibitive cost response is that it does not constitute an independent argument in favor

[5] The other major flaw, also one of labeling, is that they try to argue that their view is consistent with the basic commitments of libertarianism. It is not, but this is no objection to it.

of perfectionist policies. For we must ask why the valuable activities in question are prohibitively expensive:

> [T]hey might be expensive because there is not enough demand to make mass production feasible. But this just begs the question about why there is insufficient demand. To answer this question would then require falling back on one of the…arguments already considered. The prohibitive cost argument would thus not represent an independent argument for state perfectionism. (Quong 2011, p. 92)

Finally, he objects to the experience response by posing the question of why, assuming prohibitive cost is not the problem, perfectionist policies would be needed to ensure that (most) individuals have these valuable experiences: "If prohibitive cost is not the problem, why would people not be willing to try the opera? There are a number of possible explanations: (a) prejudice, (b) an excessively conservative disposition, (c) lack of available leisure time, or (d) lack of adequate information about the activity" (Quong 2011, p. 95). He disregards (c), on the grounds that in a society of at least Rawlsian justice, it would not be an issue. He rules out a perfectionist response based on either (a) or (b), for reasons identical to those given in the objections to responses based on weakness of will and rationality. He allows that the dissemination of information about valuable activities by the State would not be paternalistic, provided that this information is presented in a neutral way that does not amount to encouragement. But he is dubious that this use of resources by the State can be justified, claiming that "[w]hether the state, in the information age, would ever have good grounds for believing that its citizens lacked access to the relevant information about valuable activities or ways of life is uncertain."[6]

There are two errors which undermine not only Quong's objections to the perfectionist responses he imagines, but his presentation of those responses themselves. The first is the blindspot he inherits from Rawls, which has already been discussed in Chap. 9 in the context of Rawls' own theory. The Rawlsian principle of Fair Equality of Opportunity is insufficient as a guide to the just distribution of opportunities. It requires only that those of equal natural talent and ambition have an equal chance of obtaining the skills and credentials required for obtaining any desirable social position. Rawls treats talent and level of autonomous effort as given quantities which are unproblematically distributed through society. This view is unacceptable, and social justice requires much more than what Rawls has in mind when he speaks of a fair distribution of opportunities. It requires that each member of society have an equal share of autonomy-freedom—an equal chance to develop his capacity of autonomy. This is the freedom to develop the ability to reason well about which

[6] The absurdity of this claim is brought out by a recent study conducted by Farleigh Dickinson University, which reveals that individuals who get their news exclusively from either FoxNews or MSNBC know *less* about current domestic and international affairs than those who *do not watch or listen to news programs at all*. The information age has made it possible for a person to spend 24 h per day consuming the purported information delivered by a variety of sources in a variety of media, and end up less informed about the world as a result. The problem is not that individuals lack *access* to genuine information. It is that sources of genuine information are increasingly crowded out by ideologically-based sources of misinformation. It is the development of the "information age" that has *exacerbated* this problem. See (Cassino et al. 2012).

of his talents to develop, which ends to apply those talents to, and how much effort to dedicate to any given chosen end. The ability to make well-reasoned, autonomous choicesdevelopment that they must be cultivated, and of these kinds cannot simply be taken for granted, as a Rawlsian theory would have. It is a fact of human development that they must be cultivated, and their cultivation requires the right kind of social environment—an environment of autonomy-freedom. This is an environment in which those developing their autonomy have frequent access *and* exposure to a wide range of valuable activities. The most effective way to secure broad access is often through subsidies, which, we will see, cannot be cast as a form of manipulation. And the way to secure widespread exposure is by fostering an openly and energetically pluralistic society, in which sampling the wide range of valuable activities available is actively and publicly encouraged. Autonomy, in its rational dimension as much as in its dimension of self-control, is an *achievement*. The perfectionist does not think his policies are necessary because individuals make irrational choices and their behavior needs to be corrected. These policies do not work to replace the judgments of autonomous individuals. The perfectionist is concerned with how society must be structured so that individuals can *become* rational and autonomous. Perfectionist policies thus have autonomy-enhancement as their goal; and as such, they cannot be cast as paternalistic. Therefore, not only does Quong's objection to his imagined perfectionist's rationality-based response fail; Quong misrepresents the sense in which the perfectionist is concerned with rationality.

Quong may be tempted at this point to concede that perfectionist policies are necessary with respect to the young, who are still in the process of developing their autonomy, but that such policies are nonetheless paternalistic when directed at those who have already done so. But this objection cannot be sustained. As we proceed through Quong's remaining objections, we will see that perfectionist policies are necessitated not only by the need to protect the freedom of individuals to develop their autonomy, but also by the need to protect the freedom of autonomous individuals to exercise their autonomy.

Quong's Rawlsian blindspot also undermines his characterization of his imagined perfectionist's experience-based response, and at least part of his objection to it. First of all, the perfectionist need not, and should not, believe that certain particular experiences are necessary for leading a good life. Rather, it is widespread access to a broad range of valuable activities that is necessary to a just and good society in which autonomous individuals are free to pursue their conceptions of a good life. Second, Quong misses the point of the perfectionist's concern with prejudice and insularity. The perfectionist does not seek to broaden access to and publicly promote valuable activities for the sake of those who do not wish to engage in them, whatever their reasons are, provided that those individuals have had the opportunity to arrive at their preferences autonomously. He is concerned with the very real effect of the forces of prejudice and insularity on autonomous individuals. The freedom of these individuals to exercise their autonomy is limited by their social and communal environments. One goal of perfectionist policies is to help the individual overcome these limitations.

Prejudiced and insular thinking are real elements of human social life—just as real as weakness of will. And they can have a very real effect on the willingness of an individual who belongs to an insular community to seek out many of the experiences he has access to. This is true even if we assume that the individual in question has successfully developed his own autonomy, and now autonomously judges that it is in his own interest to enrich his life by having these experiences. We are concerned here with the extent of an autonomous agent's freedom to exercise his autonomy. Those who live within such communities are continuously presented with a significant obstacle to performing the sorts of "experiments in living" through which autonomy is exercised. They are discouraged from taking advantage of their autonomy-freedom in a way that others are not, and this constitutes a social injustice which requires a remedy. This is just the familiar critique of Rawls' limitation of liberal theory to the basic structure of society, pressed by G. A. Cohen, which we have already discussed, in light of which my theory of justice as Equal Liberty is constructed. Quong's puzzlement over what social justice could possibly require beyond a Rawlsian distribution of resources is a testament to the lasting importance of this critique.

We need not pretend that the effects of prejudice and insularity do not exist in order to avoid implementing paternalistic policies. A perfectionist State that broadens access to, and publicly and actively encourages its citizens to engage with, many of a broad range of valuable experiences may be motivated by an (entirely accurate) recognition of the effects of prejudice and insularity. But its actions are not paternalistic unless they diminish the autonomy of its citizens. We have already debunked the idea that subsidizing widespread access to the arts, to use Quong's preferred policy example, is a form of manipulation that diminishes individual autonomy. It should be immediately clear that there is nothing necessarily manipulative about the State promoting and encouraging attendance at cultural events—events which really do enrich the lives and expand the horizons of those who attend them. The existence of such promotion and encouragement is made necessary by the need to oppose the forces of prejudice and insularity on behalf of the individual whose life is limited by them. A State which recognizes the existence of prejudice and insularity and then responds to it in this way is making it socially easier for individuals to exercise their autonomy. Such action on the part of the State is necessarily non-paternalistic, given any reasonable understanding of paternalism. Just as perfectionist policies are needed to protect the freedom to develop one's autonomy, they are often needed to protect the freedom to exercise it. There is nothing to be gained by acting as if this is not the case.

Quong might make another concession at this point, and allow that perfectionist policies are needed to protect the freedom of individuals to exercise their autonomy, *where this freedom is actually limited by communal prejudice and insularity*. Yet he could still insist that outside of that context, perfectionist policies are necessarily paternalistic. But even this limited anti-perfectionism is unsustainable. To see why, we must discuss Quong's second error. This is his adherence, which he may not even be aware of, to the mythology of neoclassical economics. This error undermines

his objection to the prohibitive cost response, and what is left of his objection to the experience-based response.

What exactly do I mean by the mythology of neoclassical economics? We have already examined a great number of the defects of neoclassical economic theory in Chap. 10. The defect that most concerns us now is the way neoclassical theory interprets the relationship between the firm and the consumer. The most concise and eloquent statement of this aspect of the neoclassical world-view that I know of is Galbraith's:

> The essence of the neoclassical system is that individuals using income derived, in the main, from their own productive activities express their desires by the way they distribute this income for the various goods and services available to them in markets…No judgment is passed on the desires of the individual; their source is not much examined…The foregoing expression of the individual's will is passed on by the market to the produces along with the similar expressions of others. When the desire is strong, so will be the willingness to spend money. And so will be the price in the market…The producer is motivated, for the purposes of the neoclassical model exclusively, by the prospect of profit….Price changes signal to this motive…[Producers] respond. In such response they ensure that production is ultimately at the command of the individual…[T]he moral sanction of the system depends profoundly on the source of the instruction. This comes from the individual. Thus the economic system places the individual – the consumer – in ultimate command of itself. (Galbraith 1973, pp. 28–29)

The deficiency of this view should be familiar from Chap. 10. It completely ignores the fact that large corporations exert a significant amount of control over the environment in which they operate: the range of options from which the consumer can choose, the incentives or disincentives to make one choice or another, and the evidence and arguments offered (or withheld) concerning the values of the outcomes of those choices.

In the case of valuable (from the perfectionist's perspective) goods which are capable of mass production, or services which are capable of being delivered to large groups of people, Quong suggests that the perfectionist might argue that his policies are necessary to avoid prohibitive cost. The explanation for prohibitive cost that Quong offers is insufficient demand. And the only explanations he can imagine the perfectionist giving for this insufficiency of demand are irrationality, weakness of will, prejudice, insularity, or lack of information.[7] In making this argument, Quong betrays an unquestioning adherence to the mythology of neoclassical economics from the very start. As before, we must begin by reframing the way he poses the issue itself. The problem is not that the services of many large-scale artistic and

[7] We can very quickly add, without departing from the neoclassical model, one fairly obvious explanation to Quong's list. Demand for some good—say an opera company—may be insufficient within a given geographical area simply because that area is insufficiently populous to support an opera company—or any other large-scale cultural institution—if that institution must be owned and operated within the private sector, and thus generate sufficient earnings to keep its investors. The residents of such an area might all wish fervently to attend the opera, and yet the cost could be prohibitive because doing so would only be possible were they to travel many miles and incur the costs of lodging, meals, etc., away from home. It is hard to see how Quong could argue that perfectionist policies that brought the arts to these regions would be paternalistic.

cultural institutions would be prohibitively expensive if their cost were not offset by public funding. The problem is that without public support, many of these institutions would not exist. And the explanation for that is *not* that there is insufficient demand for the services they provide. The explanation is that it often does not serve the interests of any participants in the private sector to provide these services, regardless of demand.

Let us take the example of opera; similar points could be made about ballet, the symphony, etc. A full-time professional opera company is, by necessity, a fairly complex entity. The art form requires, first of all, an indoor theatre with stage, orchestra pit, and enough backstage space for dressing rooms, prop rooms, costume rooms, carpentry workshops, etc. It requires a large number of people, performing a diverse array of tasks, all of which require special skills and training: musicians and conductors; singers; choreographers; voice and accent coaches; set, costume, and lighting designers; stage managers and assistants; dramaturges; carpenters, electricians and painters; seamstresses; a sales office for advance subscriptions, an accounting department, and a marketing department; and so on. It requires a variety of less skilled workers as well: ticket sellers and takers, ushers, concession-stand operators, janitorial staff, etc. It is not the sort of endeavor which the small entrepreneur, sensing a demand in the marketplace and wishing to respond to it for the sake of profit, is in any position to embark upon.

If demand for professional opera performances in a given region were sufficient—sufficient in the sense of having the potential to generate a stable profit at a non-prohibitive price—then why would we not see large, for-profit organizations, such as corporations within the entertainment industry, creating opera companies? The answer is that an opera company, or any of a number of other types of large-scale cultural institution, is not the sort of venture which is conducive to satisfying the purposes of the technostructure. In particular, it is not the sort of venture which is capable of satisfying the affirmative purpose. The possibility of growth is severely limited by the very nature of the enterprise. Even the largest companies can only put on so many performances of so many productions in a given year. There are a fixed number of seats in a theater, and physical expansion may be extremely costly, difficult, or even impossible depending on the urban setting. If a company in a given city is successful, and more individuals would like to attend performance than can be accommodated, it is hardly the case that another company can simply be started to satisfy the excess demand. Professional-caliber orchestra musicians and opera singers are permanently scarce, given the scarcity of the immense degree of natural talent required to become one. And unlike professional sports teams—which also require complex organization, extensive personnel and infrastructure, and highly talented performers; and which are also limited by fixed schedules and seating capacity—the work of cultural organizations is not conducive to lucrative broadcast licensing and merchandising, which are the two elements of the professional sports business model with the greatest growth potential, as well as being major sources of

revenue.[8] It is also worth noting that professional sports is by no means an unaided product of the private sector. Most professional sports teams have received publicly funded subsidies for stadium construction, often in the form of local government bonds which are exempt from federal income tax. Subsidies for stadium construction since the year 2000 have averaged around 60 %. Municipal governments often also contribute extensively to all stages of planning the construction of a new stadium, and support the promotion and attendance of sporting events in a variety of ways. And the usual argument in favor of these practices—that sports franchises, and new stadiums in particular, spur local economic growth—is not supported by the evidence (Coates and Humphreys 2008).

Opera companies, and other large-scale cultural institutions, are simply not the sort of enterprises that are likely to be created by the corporate entities that would be organizationally capable of creating them. They do not serve the central purposes of those individuals to whom the decision to create them would fall—the members of the technostructure. This is so even on the assumption that sufficient demand for them exists. If this demand is to be met, there are two options remaining. One is for these cultural institutions to be founded by private partnerships formed by individuals who are both extremely wealthy—since the costs associated with initiating such a venture are bound to be large—and possess or have access to the necessary organizational capacities. This is in fact how some of the great American opera houses and companies began. The founders of New York City's Metropolitan Opera included the Morgans, the Roosevelts, and the Vanderbilts. Whether the demand for professional-caliber arts and culture is met in this way, however, depends far more on social and cultural factors than it does on economic ones.

The creation of a large cultural institution as a private partnership is hardly likely to be the *most* profitable venture such individuals could engage in with the resources they devote to it—especially given the severe limits on the possibility of growth. (This is why we may disregard as unfeasible the possibility of one being founded as a new, independent and self-contained corporation with a large number of small investors interested only in the prospect of profit.) In the case of the Met, it is well known that the motivation for its founding was the fact that the *nouveaux riches* New York industrialists were excluded by the old families of New York Society—those who were descended from the first Dutch and English families to immigrate—from subscribing to the New York Academy of Music. The founders of the Met were motivated not by the prospect of great profit, but by their appreciation for the art form and, probably most significantly, their desire for social and cultural legitimacy. Again, a comparison with professional sports is apt. The ownership model of the professional sports franchise is still that of a private partnership with a single extremely wealthy majority owner. In many cases, the franchise is far from being the most productive use these individuals could put their money to; many teams have a slim profit margin, and in 2010–2011, 33 % of professional American sports

[8] Only 50 % of New York Yankees and New York Knicks revenue, and 35 % of New York Giants revenue, is derived from ticket sales (Badenhausen 2013).

teams *lost* money.[9] The forces that make sports franchise ownership desirable are also largely social and cultural.

The private partnership model of the large cultural institution is only viable so long as one requirement is met: cultural conditions must be such that the institution continues to pay enough in "social dividends" to makes up for fiscal underperformance relative to other investment opportunities. Thus, in 1940, as the U.S. emerged from the Great Depression, and in a cultural climate tremendously different from that of the 1880s, the Met passed out of private ownership and became an independent non-profit arts and culture organization. The social dividends of ownership were no longer sufficient to justify withholding resources from more profitable opportunities, particularly given the loses suffered over the decade of the 1930s. It may have very well been the case that demand remained sufficient for the institution to turn a small profit at an accessible ticket price. But the marketplace was now caught in a dilemma. Those capable of funding the institution no longer received sufficient social dividends to justify bypassing the more lucrative opportunities their position and power gave them access to. And the small entrepreneur, who lacked access to those more lucrative opportunities and would have been satisfied with the profit margin the institution might have generated, would also have lacked the necessary organizational capacities. And yet there may well have been a sufficient audience prepared to patronize the opera at a profitable price.

This is how we came to live in an age in which the very existence of large cultural institutions relies on large non-profit arts organizations (made up of many individuals who collectively do possess the organizational capacities necessary to operate such an institution, though not to bring one into existence entirely on their own). New institutions of this type are created—planned, constructed, and launched—with the assistance of local governments.[10] And all are funded in part by sales, in part by supporter donations, but also in part by public grants—the result of perfectionist policies. These organizations satisfy a demand that the private sector is either unwilling or unable to satisfy, because the private sector does not operate in the seamless way the neoclassical model describes. One might object that if demand is really high enough that these institutions *could* be run profitably as private enterprises (if only any part of the private sector were able and willing to run them), then government subsidies would be unnecessary. They could operate as non-profits which rely exclusively on sales and donations.[11] This objection ignores the fact that in the case of creating a new institution, or taking over one that was privately owned and from which private investment has just been withdrawn, the State is often an indispensable source of funds for covering the large up-front costs any such endeavor incurs. This is the sense in which many of these institutions would not exist without perfectionist policies.

[9] As reported in *Forbes*, 7 November 2011.

[10] Precisely as in the case of new, government subsidized sports stadiums.

[11] And indeed, some do. The Metropolitan Opera currently relies on public funds for only 1 % of its annual operating budget.

The justification for *continued* State support of these institutions, still assuming they could eventually operate at non-prohibitive prices on the basis of sales and donations alone, derives from the nature of these organizations. The purposes of these organizations and the individuals who direct them are radically different from those of the large for-profit corporation and the members of the technostructure, free as they are from the need to placate shareholders and from the possibility of indefinite growth. They serve the twin goals of producing art of the highest quality, and extending it to the broadest audience they can possibly reach. The work of these organizations thus contributes directly to the expansion and exercise of autonomy-freedom, and thus to the policy goals of Equal Liberty—the appropriate goals of the State. This is what justifies continued State support of these institutions, as well as what justifies the decisions on the part of the State to contribute to the creation of one type of institution rather than another. State support allows them to contribute even more than they would otherwise be able to the achievement of the goals of the just State itself—goals which, as we will see, the citizens of a State have an exactable duty to support.

So it is not the otherwise-prohibitive cost of enjoying these institutions that necessitates their support by perfectionist policies. It is the fact that they would not exist at all without those policies. And this is explained not by insufficient demand, but by the fact that the private sector is unlikely to create them regardless of demand. Quong's presentation of the perfectionist's concern is defective, and his objection to his own framing of that concern is irrelevant to the real issue. There is no paternalism here—simply the fact that the individual does not command the efforts of the private sector in the way Quong and so many others imagine. The existence and extent of demand for arts and culture events is of course relevant to social policy decisions from a perfectionist point of view. If government subsidies make opera attendance financially feasible for every resident of a region, and all but a few make autonomous decisions not to attend, then this is a clear sign that state funds would be put to better use elsewhere. No sensible perfectionist would argue that the U.S. should have 100 professional full-time opera companies (assuming there were even enough supremely talented singers and musicians to perform in them). But this point is orthogonal to Quong's argument.

Only one part of Quong's paternalism argument remains to be refuted. He claims that although it is permissible for the State to disseminate neutral information about valuable activities and ways of life, it would be paternalistic of the State to actively promote and encourage participation in these. We have already seen that this position cannot be maintained where the freedom of individuals to exercise their autonomy is curtailed by the insularity of their communities. But it cannot be maintained outside of this context either. Such promotion does not occur in a vacuum. It occurs in the context of massive corporate efforts to shape public perception, opinion and taste in ways that further the purposes of the technostructure. Of course the effectiveness of these efforts is limited. I am not claiming that corporate persuasion leaves the consumer unable to think for himself. I do not even claim that the efforts to persuade of any single corporation, were it possible to consider their effects in isolation, would ever be found to have a *decisive* effect on any decision of any

consumer. What is important for my purposes is the cumulative effect of all corporate persuasion taken together on shaping the default value system of a culture, which cannot help but exert a significant influence on the judgments and choices of the members of that culture. I have already quoted Galbraith on this point; but it is worth doing so again: "[T]he aggregate of all such persuasion affirms in the most powerful possible manner that happiness is the result of possession and use of goods and that *pro tanto*, happiness will be enhanced in proportion as more goods are produced and consumed" (Galbraith 1973, p. 156). The point of public support for engaging in valuable activities, experiences, and ways of life is not to bend the will of an individual inhabiting a previously persuasion-free space in the way the State sees fit. The point is offer a countervailing force to the one exerted by the aggregate of corporate persuasion—to provide powerful and far-reaching testimony about what kinds of ends are valuable and choiceworthy that is not motivated by pecuniary interest. To maintain that it is paternalistic to think this countervailing force is necessary, one would have to assume that the average individual was perfectly capable of preventing this aggregate of corporate persuasion from having any effect on his judgments or decisions which he did not autonomously endorse. Here, I fear we are in the realm of fairy-tales.

Lest the appropriateness of my critique of Quong's paternalism argument be questioned, let me address a concern about the way I have understood the notion of paternalism and the basis for its being morally wrong. Quong explains his moral objection to paternalism as follows:

> I believe the presumptive wrongness of paternalism is not to be found in terms of some harm or damage to the paternalizee's interests or autonomy, but instead is to be found in a particular conception of moral status. Liberal political philosophy ought to begin with a moral or at least a political conception of ourselves as free and equal. Following Rawls, we can characterize citizens as free and equal in virtue of their possession of two moral powers: 'a capacity for a sense of justice and a capacity for a conception of the good. A sense of justice is the capacity to understand, to apply, and to act from the public conception of justice…the capacity for a conception of the good is the capacity to form, to revise, and rationally to pursue a conception of one's own rational advantage or good.' (Quong 2011, p. 100)

The first point to note is that Quong does understand the wrongness of paternalism in terms of the harm it inflicts in the form of restriction to autonomy, *on my conception of autonomy*. The capacities Rawls speaks of, the exercise of which paternalism restricts, are a close enough match for the aspects of the rational dimension of autonomy, as I have explicated these. What is really instructive about this passage is that it reveals the assumption, inherited from Rawls, which is the source of the errors that undermine Quong's anti-perfectionist arguments. This assumption, which we have already encountered, is the fundamental assumption of what Frank H. Knight called "liberal individualism," and it is the one that underlies the Rawlsian theory of social justice, among many others. It is worth revisiting the relevant passage from Knight's work:

> These reflections naturally lead up to the most important single defect, amounting to a fallacy, in liberal individualism as a social philosophy. The most general and essential fact that

> makes such a position untenable is that *liberalism takes the individual as given*, and views the social problem as one of right relations between given individuals. This is its fundamental error. The assumption that this can be done runs counter to clear and unalterable facts of life. The individual cannot be a datum for the purposes of social policy, because he is largely formed in and by the social process, and the nature of the individual must be affected by any social action. Consequently, social policy must be judged by the kind of individuals that are produced by or under it, and not merely by the type of relations which subsist among individuals taken as they stand. (Knight 1947, p. 84)

We simply cannot begin with the assumption that individuals are autonomous, or that they are free and equal, or that they have the capacities Rawls prizes, and then go from there—not if we want a system of political philosophy and a program of social policy that is at all relevant to the world we actually live in. What a theory of political and social justice must begin with is an understanding of the fact that freedom and autonomy are individual and social achievements, and that its task is to reflect on the ways social and political institutions could be structured with goal of producing individuals who are autonomous and free.

5 Raz's Perfectionism and the Contingency/Efficiency Problem

One serious problem plagues Raz's defense of his autonomy-based interpretation of the Harm Principle. The background theory of autonomy, in the context of which that defense is made, includes the following two claims: first, that autonomy does not require the presence of bad options, and respect for/promotion of autonomy does not require the preservation of such options; and second, that autonomy is not valuable in itself, but is only valuable as part of a life of self-realization. Given these two claims, it is not at all clear that Raz can maintain that it is always wrong to coerce someone who has not violated any of his duties of autonomy (and thus has not harmed anyone else). Suppose someone has chosen to reject a life of self-realization, but without doing anything that significantly limits or frustrates the autonomy of others. We may use the examples from Mill—of the idler, the drunk, and the gambler—and assume that these individuals are independently wealthy, free of dependents, fairly reclusive, and fulfill their autonomy-based duties by paying their taxes to a just State and refraining from interfering in the lives of others. The first question that confronts Raz's theory is why the State should not coerce these individuals into ceasing to make their poor choices—coerce them into not drinking so much, gambling so much, and idling the days away—assuming that this coercion can be effected in the right way. The State cannot eliminate their opportunity to drink, gamble, and idle, without violating the autonomy of many who are leading lives of self-realization. Raz is surely right when he claims that "vices and moral weaknesses are logically inseparable from the conditions of a human life that can have any moral merit" (Raz 1986, p. 168). But what if it were possible to coerce only the drunk into no longer drinking (or no longer drinking so much); to coerce

only the gambler into less frequent attendance at the gaming table; and to do so without restricting the autonomy of anyone else? Raz claims that "a moral theory which values autonomy highly can justify restricting the autonomy of one person for the sake of the autonomy of others or even of himself in the future…But it will not tolerate coercion for other reasons" (Raz 1986, pp. 173–174). But the "high value" Raz's theory gives to autonomy is strictly conditional: if one pursues a life of self-realization, then one's autonomy is very valuable. That kind of life is much more valuable when it is autonomous than when it is not. But the cases we are considering are cases of *wasted* lives. The autonomy of these individuals is not valuable. So what reason can Raz give for why the State must respect it?

Raz has two responses to this problem. One is a Millian epistemic response. He asserts that the fact that the State considers anything to be valuable or valueless is no reason for anything. Only its being valuable or valueless is a reason. If it is likely that the government will not judge such matters correctly then it has no authority to judge them at all (Raz 1986, p 412). The problem with taking this claim as a response to the present problem is that in the cases we are considering, it is implausible that the judgment that these lives are wasted is wrong. This was the same problem Mill had in making a similar response. For these lives to be valuable, there must be some valuable capacity which they cultivate. But clearly there is not. It is a fact that we can waste our lives, and the harm principle is not supported by any position that cannot produce an argument against coercing those who choose to waste their lives.

The second response focuses on the efficiency of coercion. Raz claims that "forms of coercion…all invade autonomy, and they all, at least in this world, do it in a fairly indiscriminate way. That is, there is no practical way of ensuring that the coercion will restrict the victims' choice of repugnant options but will not interfere with their other choices" (Raz 1988, p. 173). If we coerce the drunk to change his way of life by locking him away, for example, we have done nothing to improve matters. Imprisoning him will do just as good a job, if not better, of preventing him from leading a life of self-realization as he was previously doing on his own. We now have some reason for thinking that the implementation costs of coercion in this kind of case are unjustifiably high. Let us even assume that at present all our practical means of coercion suffer from this inefficiency. It is still possible that in the future we will develop more sophisticated and subtle means. If we do, how can Raz's argument justify withholding those means from individuals who have chosen to waste their lives?

This objection is what I referred to above as the efficiency problem for Raz's argument, and what Quong refers to as the contingency problem (for reasons which will become clear in a moment). Here is Quong pressing the objection:

> Suppose technological advances have made it possible to precisely control people's preferences and impulses via a chip implanted in the brain. The degree of precision is such that we could design the chips so that the one and only effect they have is to prevent us from choosing bad options. So long as we are going to make valuable choices, the chip remains inactive,but if the chip senses we are going to make an unworthy choice, the chip prevents us from doing so. The chip would not interfere with our brain, function in any other way… Should we be trying to develop more precise methods of limiting adult citizens' freedom to

> pursue bad options? Should we be investing in research and development aimed at this goal? (Quong 2011, pp. 55, 59)

Raz does consider and respond to this objection, but his response is far from adequate. He asks "what if it became possible to coerce people to avoid immoral but harmless conduct without limiting them in any other way," and then insists that "it is an advantage of [his] argument that it does depend on contingent features of our world…I do share the reluctance of supporters of the harm principle to say that in the imagined circumstances the enforcement of harmless immorality is justified," however, "it is impossible for us to say how the change would affect the merits of the issue" (Raz 1986, p. 419). One of the relevant contingent features of our world is that we do make progress in developing subtler, more efficient, more focused means of coercively affecting behavior. It does not require such a disorienting stretch of the imagination to consider a world in which the possibility Raz considers is realized. If we believe in the Harm Principle we believe that it should guide the actions of the State even in circumstances such as this. Otherwise, it is no more than a practical requirement of the moment, which should be abandoned as soon as we have the means to implement a superior principle. Raz's admission of reluctance shows that he leans toward the former position. To support the Harm Principle from within a perfectionist framework, then, we must find a way to defend it even in this hypothetical situation.

As serious as this problem is, I want to press even further in this direction than Quong does. Though it is conceptually possible to coerce someone into not choosing one type of bad option, or even many types, without coercing him into leading a life of self-realization, it is not possible to coerce someone into not choosing *every* type of bad option without thereby coercing him into a life of self-realization. For if the coerced is left incapable of choosing any bad option—including the option of idleness—then it thereby becomes impossible for him not to choose good options. And the question I want to pose is: can Raz argue that the State ought not do this, if it is capable of doing it in a sufficiently efficient and subtle way? Of course, such an agent would not by Raz's lights be autonomous. We have already seen that on Raz's account, autonomy requires the ability to choose between embracing a life of self-realization and rejecting one. But since Raz only considers autonomy valuable in the context of a life of self-realization, what basis does he have for objecting to this sort of violation of autonomy?

The anti-perfectionist liberal has no trouble endorsing a strong interpretation of the Harm Principle. He is free to judge any violation of autonomy—any act which he judges coercive or manipulative—as a harm, so long as that violation is not needed to protect another person's non-harmful exercise of autonomy from being violated. He does so not because he thinks that all exercises of autonomy are valuable, but because he insists that the State must abstain from taking a stance on which ones are valuable, by practicing either intention-neutrality or argument-neutrality. The perfectionist liberal, on the other hand, bears the burden of showing that he can endorse a plausible interpretation of the Harm Principle—one consistent with the spirit of liberalism—without abandoning his commitment to political perfectionism.

This is the last objection to the very idea of perfectionist liberalism that remains from Chap. 9. I do not think that Raz's version of perfectionist liberalism has the resources to respond to this objection. But mine does. In the rest of this chapter, I will argue for a way of understanding the value of liberty which is more complex than Raz's simple conditional account of the value of autonomy; I will develop the liberty-based interpretation of the Harm Principle required by my view, and argue for its plausibility by showing that it prohibits efficient coercion; and I will show that my version of perfectionist liberalism is fully consistent with that principle. With this argument we will finally demonstrate that endorsing the Principle of Neutrality is *not* constitutive of liberalism.

6 The Value of Liberty

Thus far we have examined the classic attempt of John Stuart Mill, and the influential contemporary attempt of Joseph Raz, to identify and justify a principle that specifies a general limit on the authority of the State to intervene in the lives of individuals. Though both attempts have been found wanting, I do think that Raz's autonomy-based interpretation of the Harm Principle is on the right track. My strategy for the remainder of this chapter is as follows. As I discussed in the introduction to Part II, it is part of my background theory of well-being that a good life is, in part, a life of freedom. Though this claim will remain a background assumption and will not receive a general defense here, I will now address the issue of the nature of the value of freedom. I will defend a particular view of the sort of value freedom has against the recent and influential theory of Ian Carter. I will then discuss the nature of the value of liberty—freedom combined with autonomy in the way I have described. A clear understanding of the sort of value that liberty has is what we need in order to proceed to examine the moral ground of the State's authority and the limits which are justifiably set on the exercise of that authority. My third task, then, will be to fulfill the promise of the conclusion to the last chapter, by applying that chapter's model of how practical authority is morally justified to the task of justifying the State's pursuit of Equal Liberty. Finally, I will advance a liberty-based interpretation of the Harm Principle and argue that it can be justified in a way that avoids the problems of Raz's view.

6.1 The Value of Freedom

The view I argue for in this section is that the value of freedom is *dependent*: that is, the value of the freedom to ϕ derives from, though is not identical to, the value of ϕ-ing. The freedom to ϕ is not valuable simply insofar as it is a freedom, without this value being related in any way to the value of ϕ-ing. Carter's position, on the contrary, is that freedoms have *independent* value: value that derives from the

simple fact that they are freedoms, without any regard for what actions they are freedoms to perform. For the purposes of this section, we should take "freedom" to refer to negative freedom in Berlin's sense, rather than to my much more restrictive notion of autonomy-freedom.

Arguing for the independent value of freedom is not Carter's ultimate goal. Rather, his goal is to argue against what he calls the *specific freedom thesis* in each of its three forms:

1. *Ontological*: There is no such thing as overall freedom.
2. *Epistemic*: Overall freedom cannot be measured.
3. *Normative*: There is no point in measuring overall freedom.

Carter is interested in arguing for freedom's independent value primarily because he believes that (a) the specific freedom thesis in all of its forms is wrong; and (b) the specific freedom thesis is wrong *if and only if* freedom has independent value. Carter never states (b) explicitly, but he does come close several times. For instance, he claims that if we deny that freedoms have independent value, we cannot be interested in how much freedom a person has overall, and we cannot justify a concern for the distribution of freedom (Carter 1999, p. 73). This rather profound confusion runs through the whole of Carter's work. That it is a confusion should be evident from our work in Part II. The claim that the value of freedom is dependent is built into the social-choice theoretic approach to measuring the extent of an agent's freedom. It is, after all, the cardinality of the set of valuable options (or of similarity partitions of such options) that count toward the extent of an agent's freedom in the models we have examined. Freedoms to choose worthless options do not count at all toward that measure, and thus cannot be said to have any value as freedoms; and this is precisely to deny the claim that freedoms themselves have some value which is quite independent of the value of what they are freedoms to do. The social choice-theoretic approach, moreover, yields a true measure of the overall freedom of agents despite its built-in assumption of freedom's dependent value. It allows us to answer precisely the question of whether one agent's opportunity set offers him more, less, or just as much freedom as another agent's. And we have seen that it is perfectly intelligible to place a concern for the extent of agents' overall freedom at the heart of a theory of distributive justice, all the while working with a measure of freedom that has the dependent value assumption built in.

I therefore dismiss both the specific freedom thesis and Carter's main argument for the independent value of freedom (namely, that the existence of such value is implied by the falsity of the specific value thesis). The work of Part II already amply supports this dismissal. I turn, then, to Carter's other arguments for the independent value of freedom. These may be grouped into two categories: (1) arguments that freedoms have instrumental independent value; and (2) arguments that freedoms have constitutive independent value. I address them in this order. I will not, however, discuss every one of Carter's arguments. A number of them rely on the assumption that the specific freedom thesis is false if and only if freedom has independent value. These will be passed over in silence.

Carter has two arguments for the instrumental independent value of freedom that deserve our attention. The first is an argument from personal ignorance (Carter 1999, p. 45). He observes that we are often enough in a position in which we know that we will have additional goals in the future, but we do not yet know what specific goals they will be. Suppose that at time t, I have the freedom to pursue option a, an option which I might adopt as one of my important goals. I value this freedom. Later on, however, at $t+1$, I have chosen not to pursue a; I have decided that I am not especially interested in a after all, and have adopted other goals from within my set of available choices. Carter points out, rightly, that this does not make it the case that I was wrong, at t, to value my freedom to pursue a. But he is wrong to think that this is an argument in favor of the independent value of freedom. The dependence thesis does not claim that only freedoms to pursue goals that an agent ends up *valuing* are valuable. It claims that only freedoms to pursue valuable goals are valuable. If a is a valuable goal—which means, from the perspective of my theory, a goal which an agent might adopt as the result of a well-executed course of ends-deliberation—then the freedom to pursue a is a valuable freedom, whether or not an agent ends up actually adopting a as one of his goals. No one has time to engage in every pursuit whose value they recognize. Those who accept the dependence thesis must claim that if a is not a valuable goal, then the freedom to pursue it is not valuable—not valuable at time $t+1$, when the agent has realized that a is not valuable, and not valuable at t, when the agent does not know that a is not valuable and still counts a as a potential goal whose choiceworthiness he has yet to examine. Carter does indeed reject this claim, but all he offers in rebuttal is the counter-claim that we are unwarranted in ruling out any of an agent's available actions as worthless or not valuable (Carter 1999, p. 54). This is patently absurd. What is true is that there are cases in which no one knows whether a particular goal or action is valuable or not. To say that so long as this remains unknown, we have some reason to protect these freedoms, is not to concede that freedom has independent value. If it becomes clear at some point that one such action or goal is without value, it will still be the case that we were right to protect the freedom to pursue it up to that point, but not because the freedom has some value which is totally unrelated to the value of that which it is a freedom to do. Rather, we will have been right precisely because we did not know whether the freedom was a valuable one or not, and that itself was a very good reason for protecting the freedom, on the chance that we would discover it to be a valuable one.

Carter's second argument is that freedom has independent value as a means to bringing about the social good of progress. This argument also appeals to ignorance—social ignorance, in this case. As a society, we do not know which freedoms will prove important for achieving the goal of social progress, and so the most effective way to pursue progress is to afford everyone the greatest possible freedom (Carter 1999, p. 46).[12] The structure of this argument is similar enough to that of the

[12] Carter rightfully acknowledges that progress is not, and should not be, society's only goal, and that the restriction of some freedoms may be required by the pursuit of other legitimate social goals.

first that the same criticisms apply. Moreover, Carter completely overlooks the fact that in addition to learning, as a society and over the course of generations, that some freedoms are especially important for pursuing progress, we also learn that some freedoms are antithetical to progress. There was a time when it was the opinion of the U.S. Supreme Court that various regulations regarding working conditions were an impermissible restriction on the freedom to contract.[13] I hope it is not too optimistic to claim that we have since learned that such restrictions on freedom are in fact essential to social progress.[14]

We can use one of Carter's own examples of independent instrumental value to identify the feature of the bearers of such value which is missing in the case of freedom (Carter 1999, p. 51). Suppose I have a ticket which entitles me to $100 worth of chocolate. The value which this ticket has for me is dependent. If I love chocolate, it will be worth a lot to me (perhaps even more than a $100 bill, which I might not be able to justify spending entirely on chocolate). If I hate chocolate or am allergic to it, it will be worth nothing to me.[15] Now suppose I have a $100 bill. I can use this to buy any bundle of goods priced at $100. The value of the bill does not depend on how valuable any given bundle of goods is to me. Its value is independent of the value to me of anything I can use it to buy, even though it is only valuable instrumentally, as a means to acquiring other things.

Carter thinks that more freedom is valuable in the same way that more money is valuable, and since the value of money is independent, so is the value of freedom. His argument for this analogy is similar enough to the two arguments already discussed that we need not examine it. But it is instructive to observe that money has independent value precisely in virtue of certain features which freedom lacks. These features are in fact the defining properties of money: it is a store of value, a unit of account, and a medium of exchange. These properties are what make its value independent. If one's income is increased and there is no worthwhile acquisition to be made, one's extra money may be saved until one comes along. And when one does come along, it will not matter what it is; it will be priced, and one will be able to exchange money for it. But freedom is not like this. Every individual freedom is a freedom to perform some specific action, to pursue some specific goal, to occupy some specific state, as Carter himself acknowledges. There are no 'freedom tickets,' granting the bearer the freedom to perform one action, whatever it may be. Carter often uses the term 'non-specific value' as a synonym for independent value. Money has non-specific value because it is itself non-specific—not connected to any specific bundle of goods. And it is non-specific in virtue of its defining properties. As freedom lacks these properties, the comparison with money, *pace* Carter, gives us reason to *doubt* the claim that freedom has independent or non-specific value.

Carter's arguments for the independent instrumental value of freedom are insufficient. Let us turn, then, to his argument for freedom's independent constitutive

[13] See *Lochner v. New York* 198 U.S. 45 (1905), among others.

[14] With respect to the Supreme Court, the lesson is generally thought to have been learned a generation later, with *West Coast Hotel Co. v. Parrish* 300 U.S. 379 (1937).

[15] We ignore here the possibility of trading the ticket for something else.

value. The claim here is that having freedoms is valuable as a constituent of a good life independent of what those freedoms are freedoms to do. The argument for this claim is as follows. Exercising one's agency, determining what one will do and what one will not do, and thus taking responsibility for the course of one's life, are part of leading a good life. Carter's crucial further claim is that "the more freedom we have, the greater the sense in which we can be called agents, and thus responsible for what we do, *because* the greater the number of times that we can say 'no'" (Carter 1999, p. 58). Carter is making a basic mistake here, and once that mistake is corrected, the argument no longer supports the claim that freedom has independent value. It is true that saying "no," deciding what one will not do and refraining from doing it, is essential to the robust exercise of agency. But it is false that the more often one says "no," regardless of what one is saying "no" to, the greater the sense in which one can be said to be an agent. And it is the latter claim that Carter makes and needs for his argument.

To see this, consider a simple example. Suppose Arthur is a diplomat, and so has diplomatic license plates on his car. These plates allow him to park wherever he likes. But as it happens, the city where his embassy is located has plentiful parking. So it is rarely if ever the case he cannot find a convenient parking spot which is legal for anyone to use. Suppose that he does park in general-use parking spots whenever he does not have a pressing reason not to. By doing so, he spares his neighbors a variety of minor inconveniences. Arthur is saying "no" several times a day. But it is just not plausible that doing so counts as an exercise of his agency of *any* significance. There is nothing valuable about the option he is saying "no" to, no reason for choosing it, no temptation to choose it. His agency would not suffer in the least if diplomats in the country he works in were to suddenly lose their special parking privileges.

Carter's assertion, then, that any freedom enhances agency in a valuable way just in virtue of the fact that it gives the agent a chance to say "no," is not a plausible one. This is enough to cast doubt on the claim that freedom has independent constitutive value, at least in the absence of another suggestion as to what such value might result from. But there is, we have noted, a related and much more plausible claim: that deciding what one will not do and refraining from doing it are essential to the robust exercise of agency. Is this claim consistent with the dependence thesis? Two considerations show that it is. First, as I have already mentioned, it is a mundane fact of life that we do not have the time and the energy that would be required to pursue every valuable endeavor that we might conceivably pursue, even if we did have the resources and abilities need to pursue each and every one of them. Operating strictly within the range of freedoms to pursue valuable goals, we have ample opportunity to decide not to engage in certain activities and then to refrain from engaging in them. We have ample opportunity, that is, to exercise our agency and take responsibility for our lives in just the way Carter is concerned with. If this were not the case, then there might be some value in the simple act of saying "no" to some options regardless of what they were. But there is simply too much in this world that is worth doing for our agency to be so easily threatened. Second, it is perfectly reasonable to count acts of moderation, temperance, and self-restraint as

part of a good life. So there is value in having the freedom to act in these ways. But such acts add value to our lives in the context of refraining from overindulging in good things, or devoting ourselves to some worthy pursuits to the exclusion of others which would add much needed balance to our lives. If I never have an opportunity to help orchestrate a genocide (to take a fairly extreme example), and so never have the opportunity to refrain from helping to orchestrate a genocide, this does not detract from the value of my life in the slightest. If anything, my life is better for never having included such a repugnant opportunity (hence the choice of such an extreme example). Far from having some independent value, some freedoms have nothing but dependent *dis*value.[16]

6.2 The Value of Liberty

I have argued, against Carter, that freedom does not have independent value. This conclusion is an important one for the discussion of the ground and limits of the State's authority that is to follow. If freedom, in the broad sense of negative freedom as understood by Berlin, does not have independent value, than the mere fact that social, political and legal institutions which are organized to promote equality of liberty (in my specific sense of this term) restrict some negative freedoms is not *in itself* a strike against such institutions. We must always look to the specific freedoms that have been restricted in order to determine whether there is any ground for objection. This section will focus on the distinction between instrumental and non-instrumental value, which clearly cuts across the distinction between dependent and independent value. I will briefly consider the question of whether freedom, in my sense of autonomy-freedom, has merely instrumental value. But my primary concern will be the nature of the value of liberty, as I have articulated this notion.

Richard Arneson has argued that the value of freedom (in Sen's sense, and thus in a sense which is sufficiently close to mine) is instrumental:

> [F]reedom is an instrumental, not a fundamental [i.e. final] value… If I know for certain that provision of opportunities would be pointlesss or counterproductive, then any moral obligation I might be under to provide those opportunities lapses. This means that even in the normal case where provision of opportunities raises the expectation that the beneficiary will put the opportunities to good use, the opportunities and resultant freedoms are properly regarded as means to a further goal, morally significant not for their own sakes but as means to individual good. (Arneson 1998, p. 192)

I have already argued, in Chap. 9, that a liberty-centered approach to distributive justice need not, as Arneson seems to think, treat opportunities themselves as final

[16] What about the opportunity to help *stop* a genocide? The freedom to do this would certainly be very valuable, as would doing it successfully, *assuming that* there was a genocide taking place. In such circumstances, it might in fact be impossible to lead a good life without having the freedom to help stop the genocide and taking advantage of that freedom. These claims are consistent with the fact that it *would have been better*, for the agent as well as humanity as a whole, if the genocide had never taken place, and thus if this achievement and the freedom to achieve it had never existed.

goods, or see each and every opportunity as a valuable constituent of every (or any) good life. Arneson now claims that to deny that freedoms and opportunities have final value *implies* that their value is merely instrumental. But there is no reason to think this. Freedoms and opportunities do of course have instrumental value. But having a suitably broad range of valuable freedoms and opportunities may also have *constitutive* value: this may be an integral part of a good life, and it may be that two lives equal in achieved functioning may fail to be equally good, because only in one—the better one—was that functioning achieved through the agent's choice of goals from a broad range of valuable options. This is, of course, part of the background theory of well-being which I have been working from, and Arneson fails to give us any reason to doubt it. This view is perfectly consistent with the (eminently plausible) claim that more freedom, even more valuable freedom, does not always make a life better, and thus that we often lack any obligation, or even any reason, to procure a given opportunity for an individual (as Arneson observes). We have already seen that a liberty-centered approach to distributive justice need not deny these claims. We should also recognize that in accepting them, such an approach does not forfeit its right to claim the freedoms and opportunities have more than just instrumental value.

I have looked at Arneson's reason for thinking that freedoms and opportunities have merely instrumental value primarily because it links up in interesting ways with various other claims of his which I have already argued against. I will refrain from further examining the reasons why someone might take freedom to have merely instrumental value. Nor will I offer a positive argument for the constitutive value of freedom; again, this is one plank of my background theory of well-being, and it will remain in the background. It is, however, important for my discussion of the authority of the State to discuss the nature of the value of *liberty*, as I have developed this notion. It is to this question that I turn for the remainder of this section.

The argument for the constitutive value of liberty is already familiar to us. It is essentially the argument that Carter used in his (failed) attempt to establish the independent constitutive value of freedom. I will not address the question of whether liberty has independent value in addition to dependent value. This is mainly because I do not see how such a project would proceed. The sort of freedom which is partly constitutive of liberty does not include the freedom to pursue worthless options. So if autonomy-freedom does have some value which is totally unrelated to the value of the options which one is free to choose, it seems impossible to tease apart these two types of value in practice. There may be some good reason to insist nonetheless that autonomy-freedom does have both these types of value. But this is a search to which I do not wish to apply myself. I will assume, then, that the value of autonomy-freedom is only dependent and that it is at least instrumental. Likewise, I will assume that the value of exercising one's autonomy is only dependent; its value derives from the value of what is autonomously chosen. If one chooses a bad option because one has deliberated badly—not in the sense of sloppily, but rather in the sense that one has knowingly discounted strong considerations against one's choice and exaggerated weak considerations for it—despite having the capacity and the access to evidence required to deliberate well, no value attaches to one's choice

merely in virtue of its being a deliberate one. Such a deliberate choice may in fact be worse than a bad choice made through sloppy deliberation or one made unreflectively. I also assume that its value is at least instrumental, since the autonomous choice and pursuit of a valuable goal are steps on the path to realizing that goal. Building upon these assumptions, I maintain only that liberty, the exercise of autonomy within a space of autonomy-freedom, has constitutive value as well as instrumental value.

Carter's argument rests on a consideration advanced by Thomas Hurka (1987). Hurka's idea, as we have seen, is that the fact that an agent is free to choose which goal to pursue from a range of options, and is responsible for the life he goes on to lead as a result of his choices *in virtue of* having chosen it, is the sort of fact that can add value to a life. A self-determined life, a life which is the product of one's agency and the exercise of one's reason, a life for which one can take credit and for which one must take responsibility, is, *ceteris paribus*, a more valuable and worthwhile life. It is instructive that the title of Hurka's piece from which Carter draws is entitled "Why Value Autonomy?" rather than "Why Value Freedom?" Hurka's conception of acting autonomously, particularly what he refers to as the Aristotelian goal of "deliberated autonomy," is strikingly close to my own neo-Aristotelian conception of exercising one's liberty—that is, acting autonomously within a suitably spacious realm of autonomy-freedom (Hurka 1993, p. 151). I may remain agnostic, then, on the issue of whether freedom by itself—even autonomy-freedom—has constitutive value. Autonomy-freedom in isolation would simply be the capability-freedom to choose one's goals autonomously *provided that* one *happens* to be able to make autonomous choices. This is a sort of freedom, that is, that can perfectly well be had both by those who can make autonomous choices and by those who cannot. Whether it is constitutively valuable or not will likely turn on whether capabilities to achieve functionings are constitutively valuable or not. I believe that they are, but I will not argue the point. I am inclined to think that the exercise of autonomy in the absence of autonomy-freedom—i.e. the autonomous choice of an option that is accessible from a set whose other members are options that merely appear to be accessible—is constitutively valuable. But since the development of autonomy requires a good deal of autonomy-freedom, the question of whether autonomous choice in isolation has constitutive value is of little practical interest or importance. It is the constitutive value of liberty that is important for our examination of the authority of the State.

7 Equal Liberty and the Authority of the State

7.1 The Moral Ground of State Authority

We can now return to the question with which we concluded Chap. 15. We want to know whether the model developed in that chapter for justifying the State's exercise of its authority can be used to justify the State's pursuit of policies of Equal Liberty. I propose to do two things in this section. The first is to show that based on the model of Chap. 15, the State has a duty to each and every one of its citizens to create the conditions of Equal Liberty, so far as it is able and consistent with a plausible limit which I will outline. The second is to show that the State has a right to compliance with Equal Liberty policies from its citizens, again within a certain limit. As we will see in the next section, that limit will turn out to be a liberty-based version of the Harm Principle.

With respect to the first task, we should begin by recalling the definition from Chap. 14 of an important interest. Such interests fall into one of four categories:

(1) An interest in the bare necessities for leading any worthwhile life which the interest-holder cannot satisfy for himself, whether by his own fault or not.
(2) An interest in being free to choose which projects and goals will structure one's life, and in being free and encouraged to develop the ability to make those choices autonomously.
(3) An interest in being free to pursue the constituents of one's well-being.
(4) An interest (a) whose satisfaction is required for the interest-holder's (continued) pursuit of his central projects and goals, (b) which the interest-holder cannot satisfy for himself without significant sacrifice and (c) for which the interest-holder's inability to satisfy it is not his fault.

These four species of interests dovetail exactly with the policies behind the Equal Liberty distribution. The first is satisfied by the guarantee of equal basic functioning, a guarantee that does not depend on whether one is responsible for one's failure to achieve that level of functioning. The second is satisfied by equal promotion of autonomy development, and equal opportunity for capability development. The goal of these policies just is to secure each person's freedom to determine his own valuable projects and goals along with the ability to exercise that freedom by making autonomous choices. The third is satisfied by the policies of comparable capability development subject to effort and capability choice, and equal freedom for capability exercise. The goal of these policies is to secure each person's freedom—in the robust sense of capability-freedom—to pursue his valuable goals. To possess this freedom is to have access to the resources required to pursue those goals, the ability (developed through one's own effort) to use those resources in order to pursue those goals, and an environment of fair competition in which to pursue those goals if they are competitive. Finally, the fourth is satisfied by the policy of leximining achievement of functioning subject to effort, accepted risk, and fair competition.

These policies leave room for compensation in precisely the sort of case which the fourth species of interest identifies.

The policies behind the Equal Liberty approach to distributive justice, then, are exceptionally well-suited as responses to the four species of important interests. That the State is the entity which is suited to the task of implementing and pursuing these policies should be clear enough. The development of infrastructure, the logistical coordination, the uniformity of administration, the *de facto* authority, and the impartiality toward each citizen which are required in the pursuit of any remotely plausible scheme of distributive justice all single-out the State as the entity that is fit to pursue such a scheme. Whether or not the State, considered as an agent, has an agent-relative reason to pursue Equal Liberty (or social justice more generally) for its citizens is a question I need not address. Finally, if the theory of social justice developed in Chap. 11 is convincing, then a truly flourishing society will be a society of Equal Liberty, and a society that falls short of this goal to some extent will fail to flourish by a like degree. We should conclude, then, that the State does have a duty to each and every one of its citizens to create and preserve conditions of Equal Liberty.

Do those subject to the State's authority have a *right* that the State pursue the goals of Equal Liberty? Can they justifiably *exact* this duty from the State? The answer to this question, narrowly understood, is negative. The reason is that in a properly functioning liberal constitutional representative democracy, such a right is unnecessary and any forcible act to compel the State to fulfill its duty is unjustifiable. Punishment for failing to fulfill this duty is delivered at the ballot box. But if we understand the question more broadly, we must answer in the affirmative. What the members of any society do have a moral right to, is to live under a properly functioning liberal constitutional representative democracy—the political system required for the effective pursuit of those goals, and in which an administration that fails to fulfill its duty can be peacefully removed from office. Respect for basic civil and political rights, democratic institutions and processes, and the rule of law, can be exacted, by force if necessary, from the public officials who constitute the State.

For the State to successfully pursue the goal of Equal Liberty, those subject to the State's authority will have to cooperate with its efforts and comply with its directives. If we assume (as we ought) that the proper goal of the State is to establish the conditions required for a society to flourish, and we are convinced by the Equal Liberty account of what it is for a society to flourish, then it is progress toward establishing the conditions for Equal Liberty that is required for the State to count as functioning well. If cooperation and compliance from its citizens is required for the State to make progress toward this goal (as they are), then the State, considered as an agent, has an important interest in this compliance; it is necessary to the pursuit of the State's proper goal. The State's interest in compliance clearly grounds an agent-relative reason for the members of the body politic to comply; it is *their* compliance which the State has an interest in, after all. And this compliance is necessary to the flourishing of society, in that it is required for the State's establishment of the conditions of Equal Liberty, which, so long as the Equal Liberty view is convincing, are the prerequisites for social flourishing.

Once we see, therefore, that the policies of the Equal Liberty approach are an ideal fit for the role of responding to each of the species of important interest which are capable of grounding moral duties, the duty of the State to pursue those policies, and the duty of the citizenry to cooperate in that pursuit, quickly follow. But it is presumably not any and every attempt to realize this goal on the part of the State that would fulfill its duty, nor any and every attempt that would generate a duty of compliance. It is not the mere fact that the State is pursuing this goal that makes it the case that its duty is discharged, or that a duty of compliance is generated. The goal must be pursued, at the very least, in a way that makes success more likely than failure (and at most, in a way that maximizes the chances of success). And this success, the realization of the conditions of Equal Liberty, must not be transitory. Progress toward this goal—or, if the goal is ever reached, the preservation of it—must be both stable and sustainable in the long-term if the State is to count as fulfilling its duty.

For the State to have a *right* to compliance, moreover, the actions it takes toward reaching this goal must result in more good than harm. These two requirements are related. We must consider the question of whether the actions of the State are the source of more good than harm from two perspectives. One is the perspective of the society as a whole, both synchronically and diachronically. From the synchronic perspective, the harm caused by the State's actions exceeds the good when the State's progress toward its distributive goal is unstable. This will be the case when, at a given point in time, it is unlikely that the State will be able to continue to progress toward its distributive goal (or to preserve the conditions it has established if the goal has been met) even in the short-term, given the course of action it is currently pursuing. The crisis of confidence on the part of the citizenry that is likely to result under this scenario will itself help to fulfill the expectation that the policies being pursued are unstable and hasten their failure. From the diachronic perspective, the actions of the State cause more harm than good when, despite the short-term or even medium-term stability of the outcomes of the State's policies, those policies are unsustainable in the long-term—i.e. their long-term effect is to make it even more difficult for the State to reach its distributive goal (or to preserve the conditions under which that goal is met) than it was at the time when the policies were first implemented.

For the State that is pursuing the conditions of Equal Liberty (or, having established those conditions, seeking to preserve them) to have a right to compliance from its citizenry, then, its pursuit must be both stable and sustainable. This means, at the very least, that the State's pursuit of this goal must be guided by sound economic policy, of the sort discussed in Chaps. 10 and 11. Since the right to compliance is the mark of a legitimate State, adherence to such policies in the course of seeking to establish (or preserve) the conditions of Equal Liberty is a requirement of the State's authority to pursue this goal. The other perspective from which we must consider whether the actions of the State result in more good than harm is the perspective of the individual within the society. And this brings us from the discussion of the ground of the State's authority to pursue Equal Liberty, and the general

constraints of stability and sustainability on that pursuit, to the question of what limits should be set on the State's power to interfere with the lives of individuals.

7.2 A Liberty-Based Interpretation of the Harm Principle

Despite its defects, Raz's argument for an autonomy-based interpretation of the Harm Principle moves us a significant distance in the right direction. Raz's notion of autonomy is closer to my notion of liberty than it is to my notion of autonomy; it includes the availability of a range of valuable options and the development of the capacities needed to take advantage of those options, in addition to autonomous choice from among those options. My account of liberty provides us with a far more robust and precise understanding of the value that Raz and I are both interested in; but the contours of his notion of autonomy match those of my notion of liberty well enough for us to see that it is one and the same value that interests us. Having already developed and defended my notion of liberty at length, I propose that we begin with a simple terminological shift. Like Raz, I will argue for a narrow reading of the Harm Principle, but on this narrow reading, harm will be taken to be restriction on liberty as I have explicated this notion. This gives us the version of the principle stated in the General Introduction at the beginning of the book:

The Harm Principle: The only adequate justification for state interference in individuals' lives is the prevention of harm in the form of restrictions on other individuals' freedom or autonomy.

That we all bear duties of liberty—corresponding to Raz's duties of autonomy—should be clear from the previous section. The citizens of the State that seeks to establish the conditions of Equal Liberty in a stable and sustainable way have an enforceable moral duty to comply with those of the State's directives that are aimed at progressing toward that goal. The liberty-based reading of the Harm Principle sets a limit on the sort of action that generates such a duty. The redistributive actions of the State, which are essential to realizing the conditions of Equal Liberty, must not be of such a rate or magnitude that those who are succeeding in leading lives of broad valuable functioning are left unable to continue (or excessively disincentivized from continuing) to exercise a range of valuable capabilities commensurate with their past effort in developing those capabilities. This limit is consistent with the Equal Liberty policy-goals discussed in Chap. 11, and thus leaves ample room for State action, allows for significant duties of liberty, and makes considerable progress toward Equal Liberty possible. Moreover, the sort of restraint imposed by this limit is the same as what is already necessitated both by the commitment to pursue conditions of Equal Liberty without wasting resources and by the fact that the State's pursuit of these conditions must be both stable and sustainable if it is to be legitimate. All three of these considerations—the Principle of No Resource Waste, the ground of the State's right to compliance with the duties of liberty it imposes, and the liberty-based reading of the Harm Principle—thus work in concert,

and are mutually reinforcing. The limitations on State action which make it possible for the State to pursue its distributive goal in the long-term (captured by the Principle of No Resource Waste) are also the limitations which the State must respect in order for its authority to pursue that goal to be legitimate; and by respecting these limits, the State remains within the confines of the liberty-based version of the Harm Principle as it imposes and exacts duties of liberty for the sake of progressing toward its distributive goal.

There remains, however, the looming objection to the Harm Principle, even in its liberty-based form, which is already familiar from our examination of Raz's argument: the efficiency problem. Let us assume that liberty does in fact have constitutive value—that the good life is, among other things, a life of liberty—but that this value is strictly dependent. Liberty does not require the presence of worthless or bad options, and there is no value in the choice of such options through the bad or sloppy exercise of one's autonomy. Let us also assume that liberalism can countenance coercion for the sake of enforcing duties of liberty, and focus our attention solely on the issues of coercing agents who are engaged in activities that are worthless but harmless to others into either (a) ceasing to choose those worthless options or (b) choosing to perform some valuable activity instead. And let us acknowledge that when our available methods of coercion are course—consisting in legal threats, sanctions and punishments—we are very likely to undermine the coerced agent's chances at achieving functioning, rather than enhancing them, and are also likely to disrupt without any justification the lives of those who are doing well. But now let us suppose that we acquire far more subtle and efficient means of coercion—subtle enough, let us say, that the one being coerced is not explicitly conscious of the presence of any threat or the fear of sanction or punishment.[17] The point of the thought-experiment is to see whether the liberal can succeed in repudiating coercion itself,

[17] George Sher discusses four types of actions which could conceivably be considered coercive. The first is punishment and threats of punishment—what I have in mind when I speak of 'coercion' *simpliciter* or 'coarse coercion.' The second is manipulation through social pressures that stop short of threats. It is an enhanced, perhaps even perfected, form of this that I have in mind when I speak of 'efficient coercion.' The third is incentivizing valuable options, and the fourth is creating valuable options and/or eliminating bad ones. I do not consider either of these to be forms of coercion. To incentivize valuable options—if what we mean by this is adding to the reasons that support choosing them, or making the reasons that already support choosing them easier to recognize—enhances autonomy, on my view of autonomy. The creation of good options increases freedom, and the elimination of bad ones does not decrease it. So neither of these types of action has a negative impact on liberty; this is why I do not consider them coercive. Sher believes that even threats and manipulation can enhance autonomy, if the coerced agent later comes to see the value of what he has been coerced into doing and would choose it autonomously if he had the opportunity to do so. He thus inverts one of the Millian anti-coercion arguments discussed above. I maintain that even in this scenario, it is still the case that the agent is not leading an autonomous life in the fullest sense, since he did not in fact autonomously choose to live the way of life which he now whole-heartedly accepts and identifies with; he was denied the opportunity to do so when he was coerced. And if there is any promise of defending a more robust sort of liberalism than this—one that places great value on leading a life of freedom and autonomy in the fullest sense—our goal should be to find a way to do so, rather than settling for the more limited variety of liberalism that Sher finds acceptable. See (Sher 1997, ch. 3–4).

irrespective of its effects, without denying that a life of valuable functioning is good to at least some (perhaps considerable) extent even if it is not a life of liberty.

The threat to the liberal's anti-coercion stance comes from the fact that, as Hurka observes, we "cannot plausibly treat autonomy [or liberty] as special among goods," by denying the plausible claim made above about a life that lacks liberty (Hurka 1993, p. 152). If we do *simply* see liberty as one good among others, then we are already set up to see our concerns about liberty as vulnerable to the problem of trade-offs with other goods. And at this point, unless we exaggerate the relative value of liberty, we will have to be prepared to yield our concern for liberty in the face of efficient, subtle and effective means of coercion. In this scenario, the harm caused by depriving the agent of liberty will more likely than not be outweighed by stopping him from making bad choices, or coercing him into making good ones. The suggestion which I will develop as I conclude this chapter is that this view gets things wrong at the start, and that recognizing this error does not require endorsing any implausible claims about the value of liberty. What is required, rather, is a proper understanding of the special *relationship* that liberty stands in to other potential constituents of the good life, which belies the claim that it is simply one good among others. And this is something we can recognize without taking liberty itself to be special among goods in any dubious way.

The structure of the conception of liberty which I developed over the first four chapters, set against the background theory of well-being I have adopted, makes the special relationship between liberty and other goods explicit. Recall that the full account of the good life is as follows: a life of achieving valuable functionings, autonomously chosen from a broad set of capabilities to function (i.e. chosen in a context of freedom), and pursued within the confines of one's moral duties. Whether a functioning counts as valuable for an agent depends on the position of the agent. Though not necessarily the functionings the particular agent would most prefer, the functionings which count as valuable for that agent are the ones that a similarly positioned agent—one with comparable natural abilities, subject to similar duties and responsibilities, and with similar tastes and interests—could reasonably include in his most preferred set of functionings. These functionings fall within the scope of the thesis of competitive value pluralism. Functionings which fit this characterization are to be judged equally choiceworthy from the perspective of society—though they are not so judged from the positions of the individual agents that belong to that society.

Insofar as we embrace value pluralism, we acknowledge that there is no one form of life whose adherents have a special claim to society's support. Insofar as we accept a competitive pluralism, we are required by our finite resources to make trade-offs; in distributing support to an agent engaged in one valuable form of life, we choose not to support others engaged in another valuable form of life. The scope of the thesis of competitive value pluralism, however, is limited to the valuable functionings which constitute the sets of options from which agents may make autonomous choices. It does not extend to the value of having broad freedom of choice among functionings, or of the development and exercise of autonomy or the making of autonomous choices. Liberty is not only a constitutive good, in the same

way that the myriad varieties of valuable functionings are. It is also a *complementary* good with respect to each and every one of these other goods. That is to say, the constitutive value of any functioning is enhanced when that functioning has been chosen autonomously from a broad range of valuable options. And the sort of conflict that can arise between the pursuit of different functionings—the conflict that is recognized by the Principle of Competitive Value Pluralism—does not arise between the functioning chosen and the liberty to choose it. Wherever there is valuable functioning, the value of that functioning increases in virtue of its being freely and autonomously chosen.[18]

The significance of this point about the complementary relationship between liberty and functioning is that there is a fundamental difference between trade-offs between functionings (which arise in virtue of competitive value pluralism) and trade-offs between functioning and liberty (as envisioned in the efficient coercion scenario). In the first case, the elements in the trade-off are in some way incompatible. In the second, there is no such incompatibility forcing our hand. Appreciating this point is the key to the liberal perfectionist solution to the efficiency problem.

The power behind the argument for coercion in the efficient coercion thought-experiment derives from the fact that there are some individuals who are failing to achieve valuable functioning, because they are choosing bad or worthless options. If these individuals were efficiently coerced into not choosing some of these bad options, some of them would probably be able to make good choices instead; freed of their temptations, they would turn themselves toward worthier ends. And the odds are that some would just make other bad choices instead; the drunk who finds himself suddenly incapable of pouring himself another drink cannot be expected to turn his life around on that basis alone (though he may). But if these latter individuals were coerced into not choosing any bad options, they would be bound to choose good ones. There would be no other possibility.

Let us consider the first case. What basis does the liberal perfectionist have for opposing this sort of limited negative efficient coercion? I am a Razian insofar as I believe that liberty does not require the presence of bad options, and respect for/promotion of liberty does not require the preservation of bad options. But I also agree with Raz that the existence of some bad options is conceptually inseparable from the existence of recognizably human lives of value and merit. Neither Raz nor his critics have given sufficient thought to the implications of holding both these

[18] This claim does *not* commit one to the sort of atomism, or insensitivity to the importance of one's place within a broader community or tradition for leading a good life, which is the focus on the liberal communitarian critique of deontological liberalism, a position which it often characterizes as unduly focused on the value of 'autonomy.' The reason is that the sorts of considerations which the communitarians focus on are perfectly legitimate inputs into an agent's deliberation of ends. These factors may have a significant impact, for example, on the categorical reasons for action which apply to the agent, since the presence of these reasons depends on a variety of facts about the agent's situation, among which the factors identified by the communitarians may figure prominently. All my view rejects is the idea that it is ever better to accept the expectations of one's community or tradition *unreflectively* than it is to choose to conform to them autonomously. For the liberal communitarian critique, see (Walzer 1983; Sandel 1982).

theses. Insofar as we are capable of eliminating bad options in such a way that we do not restrict the pursuit of good ones, we ought to do so. But eliminating bad options means changing the world we all live in, so that those options are no longer there for anyone to choose. A good, and perfectly mundane, example of this is fluoridating the water supply. This goes a long way to eliminating a bad option. It makes it much more difficult for anyone to pursue the worthless end of demineralizing his teeth to the point of decay. If we were to discover some even better dental remineralization agent, one which effectively made tooth decay impossible, had no harmful side-effects, and to which no one was allergic (as some are to fluoride), that would be even better. So to Quong's provocative question about whether we should be investing resources in discovering ways of eliminating bad options, we can answer in the affirmative, so long as we are talking about eliminating them in the sense just described.

When we efficiently coerce some particular individual into not choosing some bad option, however, we have not eliminated that option. We have not even eliminated it *for* that individual, since there is no such thing as having done that and no more; the option remains, and the coercion simply blocks the individual from choosing it. The drunk remains seated in front of his bottle, and retains the strength and coordination needed to pick up a bottle, lift it to his lips, pour its contents into his mouth and swallow them. The coercion simply stops him from picking up *this* bottle, because it is full of bourbon rather than water. While respect for, and promotion of, individual liberty does not require the preservation of bad options that can be eliminated, it does require the absence of just this sort of coercion with respect to bad options that cannot. For this sort of negative, limited efficient coercion discourages, by disincentivizing to a great degree, the development and exercise of autonomy. Most obviously, it discourages the development and exercise of one's power of self-control. The reluctant drinker who is coerced into never pouring himself another drink has lost the primary reason for cultivating his self-control so that he may autonomously resist the temptations of the bottle.[19] But it also discourages autonomy in its rational dimension. The enthusiastic lush has lost what would prompt him to reconsider his values, to further develop and then exercise his capacity for good ends-deliberation to arrive at the conclusion that a drunken life is not a worthwhile one after all. These observations are consistent with the claim that liberty does not require the presence of bad options. It is perfectly true that if, *per impossibile*, we lived in a world in which there were no bad options to be chosen, we could perfectly well still lead lives of liberty. In such a scenario, we would of course have no need for self-control, so our conception of liberty could not include this aspect. But all of that is irrelevant. It is an unalterable fact of our existence that the presence of some ineliminable bad options is conceptually inseparable from the existence of recognizably human lives of value. It is thus an unalterable fact that *our* correct conception of autonomy, and thus of liberty, must include the dimension of self-control.

[19] He has not of course, lost the *only* reason; it is still better that he not drink because he controls himself than that he not drink because he is coerced into not drinking. But by taking away the *need* for self-control, the coercion certainly discourages its development and exercise to a great extent.

Limited negative efficient coercion takes the place of self-control; it thus discourages the development and exercise of autonomy.

A similar argument can obviously be made against comprehensive negative efficient coercion. It discourages the same sort of valuable personal development to an even greater extent. An additional argument can be made against positive coercion. Once we venture down that path, even if we are successful and the agent does achieve a significant level of valuable functioning, and even if the agent comes to embrace the activity he was coerced into, we will have deprived him of one of the main constituents of a good life. It will never be the case that the agent chose his form of life, valuable as it may be, through the exercise of his autonomy and from a position of freedom.

The State that practices negative efficient coercion, whether comprehensive or merely limited, discourages the development and exercise of autonomy, certainly in its self-control dimension but likely also in its rational dimension, by creating a radical disincentive. Such a State thereby harms the ones it coerces; the radical disincentivization of the development and exercise of autonomy is a harm, by Raz's lights and my own. Quong, therefore, is wrong in thinking that the substance of the contingency/efficiency problem lies in the compatibility between Raz's view of autonomy and his autonomy-based interpretation of the Harm Principle, and the possibility of limited negative efficient coercion (Quong 2011, p. 55). The State that practices positive coercion ensures that the lives of the coerced will lack one of the major constituents of a good life. Whatever they achieve, they will not have achieved in a context of liberty. This is clearly a harm as well. These conclusions themselves, however, are not enough to avoid the efficiency problem. We may grant that efficient coercion always harms. But what reason do we have for thinking this harm is not outweighed? In the case of limited negative efficient coercion, the harm would seem to be offset by the fact that the coerced is made more likely to choose good options than he was before the State took action. In the case of comprehensive negative, and positive, coercion, good choices instead of bad are a guaranteed result. Why think that the trade-off will always go the liberal's way?

I suggested above that this way of viewing the situation—as one in which a trade-off between preserving liberty and promoting achievement must be made—was wrong from the start. Now we can see why. What appreciating the complementary relationship between the value of liberty and the value of achievement reveals to us is that this is never a trade-off we are forced to make—unlike the trade-offs that must be made between valuable forms of life that are in competition with one another. There is always, at least in principle, another option: the option of directing our efforts toward ensuring that the agent has access to a suitable range of genuinely valuable options, and changing the conditions in which the agent finds himself, such that we promote the excellent development and exercise of his capacity of autonomy. The likely result is not only an increased chance that the agent will end up achieving a significant level of valuable functioning, but also the preservation of the possibility that the agent will end up leading a good life in the most robust sense, a life of achievement that is also a life of liberty. Why should we think that promoting the conditions of liberty is likely to have this positive result, rather than to result in

an autonomous but idle or wasteful life? Here, another aspect of my theory of liberty becomes essential. On Raz's view an autonomous choice of valuable functioning is just as much an exercise of autonomy as an autonomous choice of idleness. But according to my theory, autonomy is an *excellence*, and the excellent exercise of autonomy cannot be separated from the value of the ends that are chosen autonomously. If the choice to waste one's life is an autonomous one, it is nonetheless the result of a *defective* exercise of one's capacity of autonomy. To contribute to the conditions of liberty is to alter the environment of the agent in such a way that it becomes more likely that he will develop and exercise his capacity of autonomy *well*. And the more conducive to this development and exercise an agent's environment is, the more likely he is to become an agent who is both free and autonomous, and the more likely he is, not only to lead, but what is better, to lead by his own choice, a life of valuable achievement.

Ultimately, then, the enhanced liberal defense of the Harm Principle rests on two claims. First, the fact that the trade-off between liberty and functioning is not one we can be forced to make by the sort of incompatibility that necessitates other trade-offs. And second, the fact that the extent of an agent's freedom, and the quality of his exercise of autonomy, are tied to the value of the ends that he chooses to pursue. The plausibility of both of these claims derives from the structure of the account of liberty that I have developed. What my account cannot do is enable me to demonstrate decisively that even successful efficient coercion is intrinsically wrong. What it does do is enable me to draw a weaker, but I believe still satisfactory, conclusion: we ought, as a society, to have a dominant preference for contributing to the conditions of Equal Liberty over employing efficient coercion. So long as we have not done all we can to create the conditions in which an agent will lead a life of freedom, autonomy, *and* achievement, contributing to the creation of those conditions is the preferred option over coercion (even of the efficient variety). It is preferred from the perspective of the agent, who receives support from society that further enables him to lead a good life without being subject to coercion. And it is preferred from the perspective of society, since by contributing to the conditions of liberty we both increase the chances that the individual will lead a life of value and preserve the possibility that the life he leads will include all of the major features of a good life—liberty in addition to virtue and achievement. To be a liberal is precisely to believe that the constitutive value of liberty is great enough to justify this preference.

And what if we judge that we have done all we can for a particular agent, and are unlikely to further increase his chances of leading a life of both liberty and achievement? We may, of course, be forced to make a trade-off *between agents*—say between one agent whose life is spent in worthless pursuits, and one who requires greater access to resources in order to develop a full range of capabilities or continue exercising the capabilities he has worked to develop. We may, justifiably, conclude at some point that we should no longer continue to expend resources on promoting the proper growth of the first agent's autonomy and should leave him to his idleness, focusing instead on the requirements of the other agent—and there will, we may safely assume, *always* be another such agent. There is nothing illiberal

about this course of action. In fact, it is precisely the right course of action if we are committed to promoting equality while adhering to the Principle of No Resource Waste. The essential point, the point which makes a truly liberal perfectionism fully committed to the Harm Principle possible, remains: in virtue of the complementary relationship between liberty and functioning, and the fact that both are major constituents of the good life, in the envisioned trade-off between liberty and efficiently coerced achievement, the preference for liberty dominates.

References

Arneson, R. 1998. Real freedom and distributive justice in freedom. In *Economics: New perspectives in normative analysis*, ed. J.-F. Laslier et al. New York: Routledge.

Badenhausen, K. 2013. New York Sports Teams by the Numbers. *Forbes*, 15 July 2013.

Carter, I. 1999. *A measure of freedom*. Oxford: Oxford University Press.

Cassino, D., P. Woolley, and K. Jenkins. 2012. What you know depends on what you watch: Current events knowledge across popular news sources. Farleigh Dickinson University. Available at http://publicmind.fdu.edu/2012/confirmed/final.pdf.

Coates, D., and B.R. Humphreys. 2008. Do economists reach a conclusion on subsidies for sports franchises, stadiums, and mega-events? *Econ Journal Watch* 5(3): 294–315.

Galbraith, J.K. 1973. *Economics and the public purpose*. New York: Pelican.

Hurka, T. 1987. Why value autonomy? *Social Theory and Practice* 13: 361–382.

Hurka, T. 1993. *Perfectionism*. Oxford: Oxford University Press.

Knight, F.H. 1947. Ethics and economic reform. In *Freedom and reform: Essays in economics and social philosophy*, 55–153. New York: Harper.

Lochner v. New York 198 U.S. 45 (1905).

Mill, J.S. 1869/1978. *On liberty*. ed. E. Rappaport. Indianapolis: Hackett.

Quong, J. 2011. *Liberalism without perfection*. Oxford: Clarendon Press.

Raz, J. 1986. *The morality of freedom*. Oxford: Clarendon Press.

Raz, J. 1988. Autonomy, toleration and the harm principle. In *Justifying toleration: Conceptual and historical perspectives*, ed. S. Mendus, 155–176. Cambridge: Cambridge University Press.

Sandel, M. 1982. *Liberalism and the limits of justice*. Cambridge: Cambridge University Press.

Sher, G. 1997. *Beyond neutrality: Perfectionism and politics*. Cambridge: Cambridge University Press.

Sherman, J. 2007. Utility, autonomy, and the harm principle. *Proceedings of the First Northwestern University conference in ethical theory and political philosophy*. Available at http://www.philosophy.northwestern.edu/community/nustep/07/Harm.pdf.

Thaler, R., and C. Sunstein. 2008. *Nudge: Improving decisions about health, wealth and happiness*. New Haven: Yale University Press.

Walzer, M. 1983. *Spheres of justice*. New York: Basic Books.

West Coast Hotel Co. v. Parrish 300 U.S. 379 (1937).

Conclusion

Impressions there may be which are fitted with links and which may catch hold on each other and render some sort of coalescence possible. - John Livingstone Lowes, The Road to Xanadu

In the General Introduction at the beginning of the book, I set out a group of commitments that characterize a political theory as both liberal and Aristotelian. We can now see all the basic components of this theory assembled:

1. A rigorous and precise theory of autonomy, understood as excellence in exercising the capacity to plan one's life through the deliberate choice of ends.
2. A measurable theory of freedom, understood as the freedom to develop and exercise one's autonomy, and to choose the path one's life will take from a broad range of valuable options.
3. An aspirational account of social justice which advocates the equal distribution of liberty, understood as the development and exercise of autonomy in the context of freedom.
4. A theory of the basis of the State's authority to achieve, maintain and preserve the goals of Equal Liberty.
5. A theory of the limits of State authority according to which State intervention is justified only insofar as it is needed to achieve, maintain and preserve the goals of Equal Liberty.

Much work remains to be done regarding the design and implementation of the policies that will move us toward greater social justice. But I hope to have framed a goal worth striving for.

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0

Index

J.A. Sherman, *Renewing Liberalism*, DOI 10.1007/978-3-319-28277-0