UNIVERSITY OF CALIFORNIA SAN DIEGO

A Reason-First Approach to Personal Autonomy

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Philosophy

by

Jonathan Knutzen

Committee in charge:

Professor David Brink, Chair
Professor Richard Arneson
Professor Gail Heyman
Professor Dana Nelkin
Professor Manuel Vargas

2020

The dissertation of Jonathan Knutzen is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

Chair

University of California San Diego

2020

DEDICATION

To my parents, Ken and Francie, with appreciation for their support and encouragement on this journey, and with deep gratitude for their love and friendship.

TABLE OF CONTENTS

List of Figures

ACKNOWLEDGEMENTS

I would like to acknowledge and thank David Brink, Dana Nelkin, Richard Arneson, and Manuel Vargas for their input at various stages of this project. I have benefitted immensely from their sharp questions and objections. Chapter 1 exists largely because of a challenge from Manuel. Chapter 2 exists largely because of a challenge from Dick. Chapters 1 and 5 take the shape they do because of critical challenges from Dana. And the substance of this dissertation (no pun intended) owes much to conversations with David in his office during the project's inception phase, as well as to subsequent critical input from him at many junctures.

I am also immensely grateful to David, Dana, Dick, and Manuel, for their guidance and support of my work over the years. I owe a special debt of gratitude to David for his supervision. David has been a patient and insightful critic, a tireless supporter, and (in a way that would be difficult to exaggerate) a conscientious and wise mentor.

Chapter 2 overlaps significantly with material published as "The Trouble with Formal Views of Autonomy" in the *Journal of Ethics and Social Philosophy* (forthcoming).

Vita

2002 Bachelors of Philosophy, Wheaton College
2007 Masters of Divinity, Yale University
2013 Masters of Philosophy, Brandeis University
2020 Doctor of Philosophy, UC San Diego

Publications

"The Trouble with Formal Views of Autonomy," *Journal of Ethics and Social Philosophy* (forthcoming)

"Intuitive Probabilities and the Limitation of Moral Imagination" (with Ryazanov, Rickless, Christenfeld, Nelkin), *Cognitive Science 42*, 38-68, 2018.

Fields of Study

Major Field: philosophy
Subfields: ethics, moral psychology

Abstract of the Dissertation

A Reason-First Approach to Personal Autonomy

by

Jonathan Knutzen

Doctor of Philosophy in Philosophy

University of California San Diego 2020

David Brink, chair

This dissertation defends an account of personal autonomy centered on the idea of responsiveness to genuine reasons and values, showing both why such an account is compelling and how it might plausibly be developed. Chapters 1-3 build the core case for this type of view; chapters 4-5 explore, respectively, the idea of self-governance and the value of autonomy.

Introduction

This dissertation explores the nature and significance of personal autonomy. At the heart of the account it develops is a conception of persons as normative agents with capacities for discerning and conforming to substantive truths about value. The account is at odds with a picture of autonomy that is deeply entrenched in the contemporary literature according to which autonomy is fundamentally about authenticity—about being such that one's attitudes, values, and choices are in a deep and significant sense one's own. The view I defend makes authenticity secondary, giving pride of place instead to the capacity of persons to track and pursue what genuinely matters. To accentuate the contrast with prevailing views, I call this a reason-first account of personal autonomy.

The reason-first account is a member of a broader family of views called normative competence or normative capacity accounts. Such views remain sparsely defended and relatively underdeveloped in the contemporary literature on personal autonomy. The parallel view about moral responsibility is more popular and far better developed. As I hope to show, the insights from the literature on moral responsibility have not been sufficiently appreciated for thinking about personal autonomy.

Here is a brief overview of the chapters that follow. Chapter 1 takes up a methodological challenge arising from the fact that there are, or appear to be, a variety of distinct concepts associated with our thinking about autonomy. Once we notice conceptual diversity, the question becomes what to do about it. A good approach, I argue, is to aim for theoretical unification, respecting conceptual diversity while seeking to integrate distinct ideas within an overarching

framework that holds the ideas together in an attractive way. As I show in subsequent chapters, the reason-first view helps us do that.

Chapter 2 motivates the need for an alternative to mainstream accounts of personal autonomy. According to these accounts, the conditions of autonomy can be spelled out entirely in terms of structural or procedural ingredients—things like reflective endorsement or non-alienation or coherence with one's other attitudes. Such views, I argue, face serious objections. However, normative capacity accounts in general, and the reason-first view in particular, may seem to require deeply illiberal political commitments and to be objectionable for that reason. I argue that, properly understood, normative capacity accounts need not conflict with liberal commitments.

Chapter 3 spells out the reason-first view. I discuss the view's central presuppositions and argue that they are relatively modest and should find wide appeal. I then develop the view in terms of the core idea of rational control or reasons-responsiveness, and I go on to explore a range of questions about how best to interpret that central idea. Finally, I show how the view offers an attractive framework for unifying our thinking about autonomy.

Chapter 4 explores the idea of self-governance as active self-management. There are a variety of ways to use the language of self-governance. The idea that interests me in this chapter is that agents must actively engage themselves in various ways to be effective agents. In particular, they must deploy their basic capacities for controlled activity to engage those parts of themselves that are not directly subject to their will. By doing this, agents expand control to a wider sphere of their lives. Self-governing activity of this kind merits attention because of the role it plays in helping us exercise and achieve autonomy.

Chapter 5 addresses questions about autonomy's value. Since there are a variety of distinct ideas at work in our thinking about autonomy, it is important to examine each idea in its own right.

This lends texture and nuance to the overall account while showing how the conception of agency at the heart of the reason-first view nevertheless supplies a kind of master value through which we can order and interpret the rest. I conclude the chapter by stepping back and seeking to characterize autonomy as an agency value related to our dignity as persons.

Chapter 1

Autonomy and the Challenge of Conceptual Pluralism

My inquiry begins with a methodological challenge. A number of philosophers have noted that the language of autonomy is associated with a range of concepts and gets used for a variety of purposes.[1] This observation should trouble anyone seeking to give an account of autonomy, for it raises the possibility that there can be no coherent notion of autonomy and confronts us with the basic problem of how to orient and constrain our theorizing.

The goal of this chapter is to respond to this methodological challenge. It has three subsidiary aims. The first is to introduce the reader to what I take to be the most central concepts associated with our thinking about autonomy, concepts that will appear repeatedly in later chapters. The second is to show that our thinking about autonomy is indeed characterized by genuine and deep conceptual diversity. The third is to argue that taking this fact seriously can actually be methodologically enriching.

Consider three possibilities:

1. *Conceptual monism*: There is a single privileged concept of autonomy. For all other notions we may associate with autonomy, they are either not part of the concept of autonomy, or they can be seen as interpretations of that concept.
2. *Unstructured pluralism*: There is no single privileged concept of autonomy. 'Autonomy' is a label for a grab bag of irreducibly distinct concepts which cannot be unified in any deeper way.

[1] Cf. Arpaly (2003: 117-130), G. Dworkin (1988: 6), Feinberg (1986: 27-51), Vargas (2006).

3. *Structured pluralism*: There is no single privileged concept of autonomy, but the several irreducibly distinct concepts associated with autonomy can be integrated in a systematic way as part of a theoretical package.

I will argue that a research program associated with structured pluralism is most promising. Structured pluralism offers us an attractive *via media* between the two alternatives, capturing what is compelling about each while avoiding their respective problems.

The chapter has four sections. Section 1 takes the reader on a brief tour of the conceptual landscape by introducing four concepts frequently associated with the label "autonomy." While this list is not meant to be exhaustive, it does capture some of the central terrain represented in contemporary thinking about autonomy. Sections 2 through 5 make the case for a research program associated with structured pluralism. Section 6 concludes by drawing an important methodological implication. One of my arguments for the view of autonomy I develop in this dissertation is that it allows us to honor genuine conceptual diversity while at the same time achieving theoretically attractive integration. That argument will have to wait for a fuller presentation of the view. In this chapter, I make the preliminary observation that taking structured pluralism seriously provides a useful resource for evaluating competing accounts of autonomy.

1. Conceptual Strands

Self-governance

The Greek word from which our English word, autonomy, descends means self-law or self-rule or self-governance.[2] Originally the term seems to have been used in the context of 5th century BCE Greek politics to mark a political status enjoyed by certain city states, very roughly, the status of independent self-rule.[3] Applied by analogy to persons, the idea would be that persons, like little sovereign polities are, or aspire to be, or have the right to be, self-governing.

There is evidence that the idea of autonomy was, at least sporadically, applied to persons already in antiquity. For example, in Sophocles' *Antigone*, the main character, Antigone, buries her brother against the orders of King Cleon. She thereby chooses to follow religious and customary law in defiance of the law of the local ruler. As punishment, Antigone must go to Hades. Commenting on this state of affairs, the chorus says that Antigone is the only mortal to descend to Hades alive, and that she does so "of her own law" (*autonomos*).[4] By choosing which law to follow—the law of religion and custom or the law of the king—Antigone, in a sense, makes her own law, or at any rate, makes one of these laws her own—and lives with the consequences. In that sense, she is a bit like a sovereign polity which sets its own rules for governing its internal affairs.

The political metaphor of self-governance is still resonant in our thinking about autonomy today and it crops up with some frequency in contemporary discussions. To give just one example, in their classic textbook on bioethics, Tom Beauchamp and James Childress (2008: 99–100) offer the following interpretive gloss on autonomy: "The autonomous individual acts freely in

[2] Autonomy derives from the Greek word, *autonomia*, which is comprised of the compounds for self (*auto*) and law (*nomos*).
[3] Cf. Dworkin (1988: 12-13), Feinberg (1986: 28). What exactly *autonomia* meant is more complicated. Paradigmatically, a polity would have been autonomous if it had its own army and city walls, did not pay tribute to another state, and regulated its own internal affairs by making its own laws. But the usage of this term shifted over time and seems to have been interpreted somewhat differently by the two dominant city states at the time, Athens and Sparta. For fascinating historical background, including the variety of shifting and contested meanings, see Figueira (1990).
[4] Cooper (2003: 2).

accordance with a self-chosen plan, analogous to the way an independent government manages its territories and establishes its policies."

Authenticity

A common thought is that for a person to be autonomous her attitudes and preferences must, in some special way, be her own.[5] Robert Noggle (1995: 57) gives expression to this idea when he writes, "whatever we think autonomy is, if one acts on an alien desire, one does not act autonomously."

The idea that an agent might be alienated or identified with her desires is especially associated with Harry Frankfurt's influential work on freedom, originally developed in "Freedom of the Will and the Concept of a Person" (1971). Frankfurt's central example involves a contrast between a willing and an unwilling addict. In each case, the addict has a desire for a particular drug. While there is a plain and obvious sense in which each addict's desire is his own, there is another sense in which each addict's identification or lack of identification with the desire makes it more or less truly his own. Elaborating this basic identificationist picture in subsequent work, Frankfurt describes a desire with which an agent is not identified as "external" to his will, as an "outlaw," and as an "alien" force (1999: 138; 2006: 10; 1999: 99, 136-137).

Inner psychological alienation of the sort explored by Frankfurt is one way of not being identified with one's desires and attitudes. There are other ways as well. For example, if someone is brainwashed or manipulated, or in some other way controlled from the outside, into having the attitudes and preferences she does, she may be quite happy with the attitudes and preferences she

[5] Cf. Ekstrom (2005), Frankfurt (2002), Friedman (2003), Noggle (1995).

has, yet the attitudes and preferences are not reflective of her in a very deep sense (Dworkin 1988, Mele 1995). Once one broadens the picture, it opens up the possibility that subtler forms of social influence could play a similar role in putting distance between an agent and the attitudes and preferences that characterize her. Just which forces put distance between an agent and her attitudes and preferences, and why they do so, is a matter of debate. Since we are all subject to a myriad of inner and outer forces, some account is needed which clarifies when some influence counts as reflecting or speaking for the agent and when it does not.

Many autonomy theorists take authenticity to be the core notion that explains what it is to be an autonomous agent. "So widespread is the commitment to this view," writes Michael Garnett (2013: 23), "that for many years the search for a correct theory of autonomy has been virtually synonymous with the search for a correct theory of the self." The basic idea is that an agent is to be especially identified with a privileged class of psychological structures, and that an account of autonomy is fundamentally an account of when those structures are operative.

Independence

Another common thought is that autonomous persons are independent in some important respect. There are two forms of independence which feature prominently in the autonomy literature. The first is the notion of outer, social independence. The second is the notion of inner, attitudinal independence.

Some philosophers have argued that autonomy consists in something like non-domination or non-subjection to foreign wills (Garnett 2013, 2014; Oshana 1999, 2006; Wolff 1970).[6] It is no accident that the trope of slavery is prevalent throughout discussions of personal autonomy.[7] The slave suffers one of the most profound sorts of personal unfreedom: domination under the will of another. But the image of slavery lends itself to being metaphorically extended to new contexts. One might be "dominated" by other sources of external unfreedom like a patriarchal husband or an employer or anyone who can wield tyrannical or arbitrary power over one. More generally, autonomy seems to be at odds with manipulation, compulsion, brainwashing and external control. These are forms of influence which subject persons to the will and whim of another. A natural thought, then, is that autonomy is incompatible with slavery and with sundry analogous conditions. Accordingly, a number of autonomy theorists have emphasized the need for an account of autonomy that privileges objective facts about the relations persons stand in one to another (Garnett 2014, Oshana 2006).

Other philosophers have emphasized internal independence—independence of mind. Sometimes the focus is on belief and judgments about reasons (Scanlon 1972, Wolff 1972, Westlund 2003, 2009); sometimes it is on a broader range of attitudes, including emotion, sensibility and feeling (Benson 2005, Mill 1859). Either way, the idea is that autonomous individuals form and sustain attitudes independently on the basis of their own appreciation of the world and not just because other people explicitly or implicitly commend the attitude.

[6] The best-known contemporary account of freedom as non-domination is Philip Pettit's (1997) republican theory. Pettit is careful to distinguish personal autonomy from social freedom as non-domination, though he suggests the latter may facilitate and, to some extent, be a presupposition of, the former (81-82).

[7] See, e.g., Christman (2009: 159-161, 168), Dworkin (1988: 29), Friedman (2003: 62, 191), Killmister (2009: 92-98), Oshana (1998: 81, 86).

The trope of slavery is now transferred to the internal milieu: persons can have problematically *slavish* or *subservient* attitudes. John Stuart Mill (1859) famously worries about the culturally enervating effects of conformity to what other people think and do—the "tyranny of custom" (131-138). In this respect, he speaks of the "servility of mankind" (78), of people who unthinkingly accept church dogma as "low, abject, servile type[s] of character" (116), and of people who are bold enough not to conform to custom as characterized by a "refusal to bend the knee" (131). The message is clear: independence of mind is a good thing and it is inconsistent with a servile or slavish cast of mind; it rules out *literal* obsequiousness and fawning, but it rules out subtler forms of inner obeisance as well—doing inappropriate homage to the opinions of others in the formation and sustaining of one's attitudes.

The most widely discussed example of such inner subservience in the contemporary autonomy literature is Tom Hill's (1991: 5-6) case of the deferential wife:

> This is a woman who is utterly devoted to serving her husband. She buys the clothes *he* prefers, invites the guests *he* wants to entertain, and makes love whenever *he* is in the mood…She does not simply defer to her husband in certain sphere as a trade-off for his deference in other spheres. On the contrary, she tends not to form her own interests, values, and ideals…No one is trampling on her rights, she says; for she is quite glad, and proud, to serve her husband as she does.

It is crucial to the case, as described, that it does not involve legal subjection or domination, as it might if the case were set in Victorian England.[8] Instead, the case is one of *voluntary* subservience. On some views, because of the voluntary character of her deference, Hill's deferential wife counts as autonomous (Dworkin 1988). So long as she is freely choosing her subservience, she can be autonomous in doing so. Indeed, one might think the fact that the housewife's autonomy remains intact is crucial to explaining how her case differs morally from that of the slave or the victim of brainwashing: her bowing and scraping to the wishes of her husband might be unfortunate, but it

[8] For the latter kind of case, see Pettit's (2014) discussion of Nora and Torvald from Ibsen's *A Doll's House*.

doesn't call for third-party protection or intervention in the way that the case of the slave and victim of brainwashing do. On other views, however, the housewife's subservience signals a defect in her autonomy even though it is self-inflicted and freely chosen (Benson 2005, Westlund 2009). On such views, autonomy requires some kind of inner, attitudinal, independence, not just external independence.

Autonomy as independence, then, means being independent from other persons in certain respects. It need not, of course, rule out any and all forms of dependence, many of which are necessary and desirable. Rather, it rules out pernicious forms of dependence in the form of social relations characterized by hierarchy and domination, or in the form of fawning or obsequious or extremely deferential casts of mind. These forms of dependence, whether outer or inner, are thought to be at odds with living a self-directed life.

Freedom & Responsibility

The operative notion of autonomy at the heart of our liberal social morality combines ideas of freedom and responsibility. At the heart of the liberal social vision is the idea of persons as dignified choosers who make their own choices about work, avocational pursuits, whom to associate with, and so on (Raz 1986). To the extent that they are free in making such choices, persons will also be responsible for the shape of their lives.

Following Isaiah Berlin, philosophers distinguish between positive and negative liberty (Berlin 1958/2002). Berlin contrasted negative liberty, which consists in freedom from interference by others, from positive liberty, which consists in self-realization or self-mastery. For present purposes, a slightly wider contrast is useful: between freedom from interference by others,

on the one hand, and an agent's abilities and powers, on the other hand, along with associated goods and opportunities which enable her to achieve her ends—adequate resources, skills, discipline, wisdom, or whatever. Autonomy is associated with both. On the one hand, respecting an individual's proper sphere of choice means treating her as presumptively sovereign in that sphere (Feinberg 1986), that is, not interfering with her choice. On the other hand, if people are to really make meaningful choices in life for which they can be responsible, they need suitable options and the agential skills and capacities to succeed in competently selecting among the options.

Consider a particular context: medical decision-making. Patients should be free to make their own decisions about their treatment. This means that, so long as they are sufficiently "capacitous," their decision may not be usurped or tampered with by others. Certain forms of influence on their choice—threat, compulsion, deception—must be avoided. However, in order for patients to make suitably free choices, there is reason to address other kinds of threats to freedom as well: to properly inform them, to calm their anxieties and help them think clearly, to ensure that their decisions are not driven by an avoidably constrained option set (e.g., through lack of resources), and so on. Patient autonomy, then, plausibly requires safeguards against interference as well as positive empowerment. The same can be said for autonomy more generally. To be an agent who makes life choices that are meaningfully free and responsible requires not only protection against external interference but also personal empowerment.[9]

[9] It is controversial whether the state may promote positive freedom and autonomy. I'll return to this issue in the next chapter. For now, note that those who argue that the state should only promote negative liberty should have no qualms accepting that *other* agents may promote positive liberty or autonomy *in a private capacity*. For example, presumably parents may promote the positive liberty or autonomy of their children through character formation and material resourcing.

The operative liberal notion of autonomy also implies the idea of responsibility.[10] It is not hard to see why. First, diminished personal freedom, and therefore diminished autonomy, will tend to diminish responsibility. The connection between freedom and responsibility is proportionate and scalar: all else equal, the more freedom one enjoys, the more responsible one tends to be for one's choices. Consider some of the external and internal freedom-impairing conditions mentioned above. For example, the slave is not responsible for the shape of her life (only, perhaps, for a tiny sliver); the person subject to crippling phobias is not responsible (or not very responsible) for the choices that issue from her condition; the person who is manipulated into doing certain things is less responsible than if she is unmanipulated, and so on. In other words, responsibility for acts and outcomes is sensitive to the extent to which agents enjoy freedom/autonomy-friendly circumstances.

Second, threshold-level competencies for autonomy go with, and seem naturally explicable in terms of, responsible agency. Consider norms against paternalism. It is generally those agents who are capable of being responsible for their own lives and choices whom it would be presumptively wrong to paternalize. That presumption is strongest above some threshold-level of minimal competence that marks out a sphere of responsible agency; it is absent in agents below that threshold, like very small children and non-human animals, who are incapable of being responsible for their lives and choices in any meaningful sense; and it is attenuated for agents, like school aged children, who have diminished or fragile responsibility capacities. Which facts determine the appropriate quality and scope of paternalistic interference for such agents? The answer must surely be sensitive to facts about their responsibility capacities. Older children will tend to merit stronger protections against paternalism and enjoy a larger sphere of protected choice

[10] Cf. Arneson (1980: 475), Buss (2012: 648), Dworkin, G. (1988: 20), Dworkin, R. (1999: 224), Friedman (2003: 21-22), Killmister (2018: 4, 135-142), Westlund (2009: 30-36).

than younger children. This is because they tend to have enhanced capacities for making choices for which they can be meaningfully held responsible. Age, not itself of direct normative relevance, is presumably a rough proxy for such capacities.

The link between autonomy and responsibility seems to be a fairly deep presupposition of liberal anti-paternalist principles more generally. It is widely agreed that across a broad swath of life choices, people should be allowed to make their own choices as they see fit—even when they choose unwisely. One of the functions of anti-paternalist norms in liberal society is to safeguard people's freedom to make self-regarding choices as they see fit—even when there is reason to suspect they will use this freedom unwisely. But why respect unwise choices? Admittedly, there are a variety of sources of justification for anti-paternalist principles. Perhaps individuals tend to know best what is in their own interest, perhaps governments and other intervening agencies cannot be trusted, and so on. But, as I will argue in the next chapter, it is difficult to conceive of a complete and adequate answer to the question why we should let people make their own decisions, often about matters of grave consequence, without appealing to the thought that persons are agents capable of being responsible for their own lives and choices. Whatever other justifications there may be for the strong anti-paternalist norms typical in liberal societies, the idea that persons are responsible agents seems to be a fairly deep presupposition of our practices. As noted, the best explanation for the difference in treatment of children and adults would seem to be a difference in capacities for responsible agency. The more we think children resemble adults in being responsible agents, the less we will think it appropriate to paternalize them, and the more we think adults resemble children in being non-responsible agents, the more we will think it appropriate to subject them to paternalistic treatment.

The notion of autonomy at work in our liberal social morality is thus evidently closely connected with the twin ideas of freedom and responsibility. We might put this by saying that autonomy is a form of responsibility-entailing freedom. To be an agent capable of autonomy is to be an agent capable of responsibility, and for such an agent to enjoy circumstances favorable to the exercise of her autonomy competencies is for her to be responsible in some significant sense for the upshots of her choice.

2. Three Interpretations

The last section introduced four autonomy concepts: self-governance, authenticity, independence, and responsibility-entailing freedom. The list could be expanded or contracted. It could be contracted by scrapping categories or reducing one category to another. I will return to this possibility below. It could be expanded by adding concepts, for example, self-direction and self-authorship. In my view, concepts like self-direction and self-authorship are not fundamental, since they can be explained in terms of concepts already on the list. Self-direction might mean having and living from an authentic self, or it might mean enjoying inner or outer independence; self-authorship might mean that one's life is deeply one's own, or that one is free and responsible, and so on. I believe the list represents a fairly good approximation to the central conceptual terrain in our thinking about personal autonomy.[11] Still, I leave the possibility of expansion open. Perhaps there are further fundamental concepts that should be added to the list.

[11] One might, of course, count differently, e.g., by counting inner and outer independence as separate concepts, or by counting freedom and responsibility, and again, positive and negative freedom, separately. What matters, however, is that we make the relevant distinctions, not how we count. My argument does not depend on counting in any particular way.

Let's turn to the question of how to interpret the appearance of conceptual diversity. Consider the three possibilities mentioned at the outset.

Conceptual monism. The first possibility is that there is a common conceptual core, which unifies the different autonomy concepts. One way to locate such a core is to take guidance from etymology. For example, in a *Stanford Encyclopedia of Philosophy* survey article, John Christman (2018) writes that "a theory of autonomy is simply a construction of a concept aimed at capturing the general sense of 'self-rule' or 'self-government.'" The most ambitious proposal of this sort is due to Joel Feinberg (1986), who claims that the idea of self-governance supplies the scaffolding for our entire thought about autonomy.

Another way to locate such a conceptual core is to look for overlap. For example, James Stacey Taylor's (2009: 2-3) suggests a "capturing analysis" in which what most people (or most philosophers) mean by autonomy is captured by a theoretical account. Recognizing that etymology is often a poor guide to contemporary meaning and that a variety of ideas not readily subsumed under the concept of self-governance are at work in our thinking about autonomy, this approach eschews etymology in favor of a more capacious conceptual core. The basic strategy for unification, however, is the same: locate a core concept.

Unstructured pluralism. Reviewing a collection of essays, all ostensibly on the topic of autonomy, Manuel Vargas (2006) writes:

> […] after reading through it one might justifiably wonder if there really is a unified field of philosophical work on autonomy. The diversity of essays in the volume makes a perhaps inadvertently compelling case that a number of distinct—and at best loosely-related—conversations share the same subject matter only in name…This is, of course, not an objection to the work of any particular author or even the volume itself. It is only to observe that if autonomy is one thing it is protean.

Nomy Arpaly (2003) raises similar worries about the unity and usefulness of autonomy talk. Neither Vargas nor Arpaly settle for any definitive interpretation. They each seem to recognize genuine agency ideals associated with the label 'autonomy,' and they leave open the possibility that at least some of these ideals are appropriately pursued under that label. Nevertheless, we can distill from their skeptical observations the following possible interpretation of the situation. Different philosophers use the label 'autonomy' to talk about different phenomena. We shouldn't think of this as a situation in which these authors are talking about the *same* thing; rather, we should think of it as a situation in which they are talking about a variety of *different* things. It turns out that there are just a variety of different concerns and projects carried out under the banner of 'autonomy.' On this interpretation, there is no deeper unity behind the appearances: disunity goes all the way down. A diversity of phenomena have been tagged with the same label. Unity is merely nominal.

Structured pluralism. Instead of looking for unity at the conceptual level, we might look for it at the level of theory construction. Here is the basic idea. There are—this is the truth in pluralism—a variety of irreducibly distinct ideas associated with our thought about autonomy. The challenge is to take these ideas and see if one can put them together into a theoretically satisfying package. This is similar to how we proceed in other areas of normative inquiry. Take the idea of social equality. There are a variety of other notions related to this idea: dignity, standing, fairness, relative well-being, absolute well-being, and so on. There is little reason to think these ideas all fall out of some conceptually unified architecture. However, it *is* plausible to think that these ideas can be theoretically unified: competing theories of equality are proposals about how to put together different elements to produce an overall coherent and sensible vision of social equality. Similarly,

we might hope that the various ideas associated with our thought about autonomy can be integrated within some more encompassing theoretical framework.

Each of these three interpretative options represents a plausible conjecture. In order to decide which of the conjectures is most plausible, we need to try to answer two basic questions. First, is there deeper unity? Second, if there is deeper unity, what is its source?

If there is *no* deeper unity, then unstructured pluralism is correct. This leaves two further possibilities. Either the language of autonomy should be dispensed with and whatever other concepts are useful should be retained (eliminitivism), or the language of autonomy should be retained but it must be recognized that there are several autonomy concepts which cannot be unified in any meaningful way (radical pluralism).

If there *is* deeper unity, then either conceptual monism or structured pluralism is correct. Which one is correct will depend on whether the source of unity is found in a core concept or not. Here is a diagram of the possibilities:

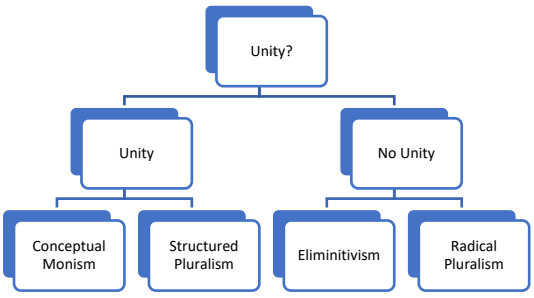

Figure 1: diagram representing the space of possibilities:

unity (conceptual monism, structured pluralism) vs. no unity (eliminitism, radical pluralism).

Let's consider each of these four options in turn, beginning with the no-unity branch.

3. No Unity

Unstructured pluralism may seem most naturally to recommend eliminitivism. If the only genuine unity is provided by a common label, it may be best to dispense with the label and concentrate instead on illuminating the underlying phenomena of interest. However, it is in principle conceivable that the language of autonomy ought to be retained for a variety of distinct concerns which cannot ultimately be unified. In the first case, autonomy would be like phlogiston: a theoretical posit which ought to be eliminated. In the second case, autonomy would be more like jade: a common label whose reference is divided.[12] Each of these options is difficult to accept.

Eliminitivism

Eliminitivism suggests we could replace the language of autonomy with a patchwork of more specific ideas without any loss. But given how entrenched the notion of autonomy appears to be, and given that it seems to do significant moral work, this is difficult to maintain. Think, for example, of medical decision-making and liberal anti-paternalist norms. On its face, it looks like the value of autonomy plays a role in each of these contexts: it is part of what explains why persons' self-regarding choices must be respected. To be sure, eliminitivism need not deny normatively important phenomena. One might, for example, accept that persons' self-regarding choices must be respected but think this is explained by their being free and responsible. Ultimately, however, eliminitivism does recommend dispensing with autonomy talk—and this is difficult to accept.

[12] Jade, it turns out, is a name for two different natural kinds, jadeite and nephrite. The example is originally due to Hilary Putnam (1975: 241).

The language of autonomy is by now a well-entrenched feature of our international moral, political, and legal culture.[13] Correspondingly, the idea of autonomy is fairly indispensable to contemporary moral thought. It is, as just noted, a central value in medical ethics. More broadly, it is a recognizable value in liberal social orders, which prize self-direction and (some form of) independence, and which are committed to protecting a significant sphere for individual choice. As noted above, at the heart of this social vision is the idea of persons as dignified choosers who must chart their own course through life. This idea marks out two roles for the idea of autonomy. One is an agency ideal: all else equal, autonomy is a desirable agency characteristic. Another is a principle protecting the exercise of this sort of agency: the choices of an autonomous agent call for respect. These are familiar ideas. The language of autonomy seems both well-suited to talking about them and not in any obvious way replaceable. Unless we think the language of autonomy is entirely misplaced, eliminitivism seems like a rather extreme proposal and will be difficult to accept.

A less drastic proposal would be to say that we should prune our thinking about autonomy, cutting some conceptual strands and retaining others. But this possibility amounts to reformism rather than eliminitivism. If there is some pruned successor concept for which we should retain the language of autonomy, eliminitivism cannot be right. For if there are presently a variety of ideas floating around under the banner of autonomy and one of these is the true or useful idea of autonomy, then the language of autonomy needs to be taken seriously and arguments must be given for thinking one way of understanding autonomy has important merits over another. The reformist program is in the end incompatible with eliminitivism, since it recommends hanging on to the language of autonomy.

[13] Cf. Feinberg (1986), Möller (2012), Raz (1986).

Radical Pluralism

It is plausible that a reformist program would need to locate the conceptual or theoretical package that preserves the best parts of our autonomy-talk and earns the keep of its continued use. This package would need to be more than merely nominally unified: it would need to hang together in some deeper way. If radical pluralism is true, however, that assumption is mistaken. On the radical pluralist interpretation, it would be true *both* that (i) we should retain the language of autonomy for different phenomena (phenomenon 1, phenomenon 2, phenomenon 3, etc.), *and* (ii) that there is no deeper unity to these phenomena, so that we must recognize fundamentally different kinds of autonomy corresponding to the different phenomena (autonomy$_1$, autonomy$_2$, autonomy$_3$, etc.).

Here is a possible analogy. Often it turns out that on closer inspection there are a variety of distinct concepts of a certain kind, yet it remains unclear what, if anything, unifies them. For example, Steven Darwall (1977) argues that respect comes in two varieties: recognition respect and appraisal respect. Similarly, Gary Watson (1996) argues that responsibility comes in two varieties: attributability and accountability. Many philosophers now accept that there are at least two kinds of respect and at least two kinds of responsibility (if not more). Might it be similar with autonomy?

The suggestion is frankly difficult to make sense of. In the cases of respect and responsibility, we say that there are two *varieties* or *kinds* or *forms* of these things. This implies deeper unity. If respect$_1$ and respect$_2$ are both *kinds* of respect, then there must be some superordinate category of which they are both members and to which they stand in a genus-species

relation. Such nested structure is at odds with radical pluralism. Similarly, if autonomy$_1$, autonomy$_2$, autonomy$_3$, and so on, are all *kinds* of autonomy, then it is hard to avoid the conclusion that there is some form of deeper unity which makes them all species of a kind. Of course, one might be reformist here and say that certain instances of so-called respect or responsibility or autonomy are not genuine kinds of respect or responsibility or autonomy. But that not radical pluralism.

How are we to understand the idea of radical pluralism? Suppose there really is no deeper unity. Then in virtue of what is it important to use the same language to cover each of the different cases? Since the various cases have all been tagged with the same label, it may be a matter of convenience to keep calling them all by the same name, but this is hardly vindication for the theoretical importance of clinging to the language. Consequently, radical pluralism is unstable: if there is no deeper unity, it is not clear why we need a common label; if a common label is needed, then it is hard to deny that there is some deeper unity in virtue of which the common label is appropriate.

The no-unity branch, then, does not seem promising. Unless we are skeptical that the language of autonomy, which is so central to our moral and political life, is radically mistaken, we have reason to accept a research program which vindicates at least some parts of that language. If the program is conservative, it will vindicate more of the existing language; if it is reformist, it will vindicate less. Either way, however, the program will seek unity in a conserved idea of autonomy.

4. Unity

There are two plausible candidates for deeper unity. Conceptual unification locates deeper unity by way of a conceptual core; nonconceptual unification locates it in a theoretical model. Let's look at each possibility.

Conceptual Monism

What might conceptual unification look like? One example is Feinberg's (1986) proposal that self-governance is the core autonomy concept. Another is Taylor's (2009: 2-3) proposal of a "capturing analysis," which aims to distill a focal meaning from what most people (or most philosophers) understand by autonomy. In Feinberg's version of conceptual unification, one of the existing autonomy concepts (self-governance) is privileged over the others and used as a sort of interpretive and organizing lens; in Taylor's version, a concept not necessarily identical to any of the special autonomy concepts is distilled through a process of analytical distillation. Either way, a core concept is envisioned as doing the unifying work.

It might be tempting to interpret conceptual unification of either kind as a *semantic* project.[14] A semantic project, however, is a poor candidate for achieving unity. Clearly, the different ideas of autonomy canvassed in section 1 do not *mean* the same thing; they mean quite different things. That is in many ways the problem. If it were transparent just by looking at the meaning of the different ideas how they relate, there would presumably be no interesting problem about conceptual diversity that would prompt us to search for deeper unity.

[14] Some things Feinberg and Taylor say encourage this interpretation. Feinberg says autonomy has four closely related *meanings* (28); Taylor speaks of his method as an *analysis* meant to capture *what people mean*, and he says that a philosophical clarification of the target concept takes as input, and must appropriately respect, the *connotative contours* of the concept (2).

More plausibly, conceptual unification might be interpreted in epistemic or metaphysical terms. On an epistemic interpretation, a concept's unifying power would consist in its ability to illuminate and make sense of other concepts. On a metaphysical interpretation, a concept's unifying power would consist in a fact about how its instantiation relates to the instantiation of other concepts. Take the concept of self-governance. According to the epistemic interpretation, the concept of self-governance helps us better understand or interpret other ideas, like independence and authenticity. According to the metaphysical interpretation, the concept of self-governance implies ideas like independence and authenticity because they are constitutively involved in what it is to be a self-governing agent (i.e., one counts as self-governing only if one has an authentic self or enjoys mental or social independence).

Conceptual unification of either sort faces a serious *prima facie* challenge. The ideas canvassed in section 1—self-governance, authenticity, inner and outer independence, and responsibility-entailing freedom—appear to be distinct and none seems to enjoy any obvious priority over the others. For example, the kind of freedom that would constitute one as responsible seems quite different from authenticity. A precocious child might have a developed sense of self and, acting from that sense of self, be quite authentic, yet lack (developed) capacities for responsibility. Conversely, one might be a free and responsible adult yet not have a very developed sense of self or fail to act in conformity with it. Both of these ideas in turn seem different from independence. Independence itself is not one thing but two, neither of which is reducible to the other: one might enjoy attitudinal independence without social independence (the Stoic slave envisioned by Epictetus) or social independence without attitudinal independence (the average high schooler). So far as I can see, there is irreducible and deep plurality here; none of the concepts maps neatly on to the others. Moreover, none of these concepts seems to deserve the epithet

'autonomy' more than any other: they are all on equal footing as far as that goes. So while there are no doubt rich and interesting connections between the concepts, it is not clear one of them is epistemically or metaphysically primary.

If one accepts that conceptual diversity is genuine, and if one does not judge any one idea to be *the real* or *central* idea of autonomy, then prospects for conceptual monism look dim. Consequently, invoking a concept like self-governance is unlikely to succeed in bringing about deep conceptual unification. For each way of spelling out self-governance that gives it substantive content, it will tend to track one of the distinct conceptual strands in our thinking about autonomy and lose the others. Suppose self-governance is understood as something like being a self-directed agent, which in this context we can understand as living from authentic preferences and values. Then self-governance (unsurprisingly) can capture the idea of authenticity, but arguably it does not capture the idea of responsible agency which is also associated with autonomy. Or suppose self-governance is something more like a socially independent self—one who is, say, undominated and unoppressed. Then self-governance (unsurprisingly) can capture the idea of social independence, but it arguably does not capture attitudinal independence or responsibility. And so on. If the appearances are genuine and conceptual diversity runs deep, then this is exactly what we should expect.

Taylor's suggestion of a capturing analysis might seem more promising, since the resulting conceptual core could in principle be quite capacious. However, this suggestion is vulnerable to a problem roughly analogous to the one we just identified about privileging the concept of self-governance. If the capturing analysis preserves genuine and deep conceptual diversity, the resulting conceptual core will end up being a mere conceptual amalgam, not a unified concept; if,

on the other hand, it privileges one idea over the others, the resulting conceptual core will end up not doing justice to conceptual diversity.

It all depends, of course, how deep conceptual diversity goes. If we judge that ideas like authenticity, independence, and responsibility-entailing freedom are all genuinely part of our thinking about autonomy, and if we think these ideas are fundamentally distinct, then a notion of autonomy that succeeds at capturing our thought needs to do justice to these diverse elements. It isn't clear a capturing analysis is the appropriate tool for this.

Conceptual monism looks more attractive when paired with a strong reformist program. Perhaps, for example, autonomy should be identified with authenticity but not with the sort of freedom that makes one responsible, or vice versa. This would preserve conceptual diversity while dividing conceptual labor: autonomy is one thing and it plays a certain conceptual and normative role, and then there are a variety of allied concepts which may be more or less loosely associated with autonomy, which play different conceptual and normative roles. In principle, such conceptual division of labor is attractive, but the case for it needs to be made. As I hope section 1 revealed, there is a strong *prima facie* case that our thinking about autonomy is characterized by genuine conceptual diversity and by a variety of normative concerns associated with different elements. So while the attraction of identifying the idea of autonomy with just one of these conceptual strands is intelligible, it needs to be shown why this is a compelling interpretation in its own right and not simply a way to avoid facing the troubling appearance of conceptual diversity. It seems to me that systematic pressures pull in different directions and that an adequate picture of autonomy needs to be faithful to this complexity. If that is right, a single-strand view would yield, not an elegant division of conceptual labor but an impoverished account of autonomy.

Conceptual monism seems attractive because it promises to avoid the troubles of the no-unity branch and do so by way of a very sensible-looking strategy. However, if I'm right, conceptual monism underestimates the amount and depth of conceptual diversity and overestimates the power of unification around a single, central concept.

Structured Pluralism

To find unity, we need not suppose that all of the apparently diverse strands in our thinking about autonomy can ultimately be understood as just so many buttresses in a *conceptually* unified architecture. Instead, integration can result from a process of theoretical construction.

Return to an analogy from earlier. Compare the project of giving an account of autonomy to the project of giving an account of social equality. The notion of social equality is highly complex; it involves a variety of other concepts like standing, fairness, relative well-being, absolute well-being, and so on. The goal of an account of social equality is to integrate the various concepts into a coherent and attractive framework which does justice to a variety of competing normative pressures. It seems unlikely that any one of these concepts is the conceptual master key that helps unlock the others. Instead, it seems likely that the way to achieve integration is to find a suitable theoretical framework to hold the various concepts together. The work of integration, if successful, occurs at the level of theory, not at the level of concept.

Similarly, the goal of an account of autonomy, it might plausibly be thought, is to develop a theoretical framework that can draw the various conceptual strands of our thinking together into an attractive whole. On this picture, rather than thinking of the resulting account as a complex conception of a single concept, "autonomy" would be the name for the solution to a theoretical

puzzle, a solution which allows us to hang on to, and make sense of, a variety of distinct concepts. As in the case of social equality, the work of integration would occur at the level of theory, not at the level of concept.

Philosophers sometimes claim that autonomy is a term of art (Dworkin 1988: 6-7, McKenna 2005: 206). This makes it sound like autonomy can be given whatever stipulative definition we like. But that misconstrues the situation. As noted above, autonomy is a recognizable ideal in liberal social orders that assumes a certain picture of individuals, of their potentialities, of what is a good way to live, and of how people ought to relate to one another. This multi-faceted picture covers a cluster of ideals, norms, and practices, including respect for individual choice and its associated anti-paternalist norms, ideals of social and mental independence, and ideals of authentic selfhood. The hypothesis of structured pluralism is that this cluster, though complex and involving a variety of distinct concepts and normative concerns, is not an entirely random or normatively disjoint assortment. In the next section, I'll suggest a concrete place to begin searching for theoretical integration.

5. The Case for Structured Pluralism

My argument for structured pluralism has been that it represents an attractive *via media* between conceptual monism and unstructured pluralism. In a nutshell, the argument is that structured pluralism gives us a way to hang on to what is attractive about the alternatives while eschewing what is unattractive about them. On the one hand, it allows us to accept that conceptual pluralism is genuine and deep. On the other hand, it gives us a way of seeking to vindicate the non-skeptical result that our autonomy-related discourse and practice has a solid and defensible core,

and that a variety of distinct inquiries carried out under the banner of autonomy are not misguided in thinking they really are about autonomy.

The argument makes a number of assumptions:

1. *Pluralism*: There are multiple genuine autonomy concepts.
2. *Irreducibility*: These concepts are not reducible one to another.
3. *Non-priority*: None of the autonomy concepts is more basic or privileged.
4. *Indispensability*: These concepts are all deeply implicated in our moral and political thinking and cannot be dispensed with.
5. *Conservatism*: The language of autonomy seems appropriate to these concepts. We should try to vindicate that use before jettisoning it.

This package of commitments speaks in favor of structured pluralism. Not everyone will accept these claims, but those who do will find that they are naturally pushed toward an alternative to no-unity views as well as to unity views which depend heavily on the assumption that there must be one, central, and privileged concept of autonomy.

6. Methodological Upshot

Many philosophers, I suspect, operate with the implicit assumption that there is a single concept of autonomy and that they are, or ought to be, in the business of pursuing conceptual analysis to elucidate that concept. This methodological approach fits naturally with the underlying assumption of conceptual monism. However, if there are several irreducibly distinct concepts, none of which is privileged, it won't do to try to find *the* concept of autonomy and then give an account of it, as conceptual monism recommends. Instead, assuming the aim of unity is desirable,

the task must be to identify the diverse concepts and then to try to systematically unify these into a coherent and attractive package. The basic research strategy associated with structured pluralism encourages us to seek unification at the level of an integrating theoretical account. Once this strategy is made explicit, it supplies an important resource for regulating inquiry. Since the strategy tells us to aim for integration, how well an account does in fulfilling this aim can supply a useful criterion of theory choice. For we can ask: how well does a candidate view of autonomy do in unifying diverse conceptual elements? All else equal, the more an account is able to unify, the better.

Let me conclude by suggesting where we might look for theoretical unity. On a plausible assumption, autonomy only becomes relevant when we are dealing with agents who have certain capacities. Whatever exactly autonomy is, it appears to be grounded in basic capacities which we associate with personhood and normative agency. For example, mature persons are the sort of creatures who are capable of forming an evaluative conception of what matters in life—a "conception of the good," in Rawls's (1971) famous phrase—and being guided by this conception. It seems to me that at the heart of our thinking about autonomy has to be some such story about personhood and normative agency. A plausible place to search for an integrating account of autonomy, therefore, is in such a story. This might provide resources to weave the various conceptual strands together into a satisfying whole. In the very broadest terms, autonomy would refer to a cluster of values and principles associated with this special form of agency of which persons are capable. It would incorporate agency ideals, principles for the protection of this valuable sort of agency, and specifications of the inner and outer milieu that are hospitable to the exercise of such agency. In short, autonomy would refer, not to a single property or value, but to

a variety of properties and normative concerns which have as their unifying rationale the exercise or fulfillment of this special form of agency.

Now there are broadly two ways to fill in the story about normative agency. The first makes no reference to substantively rational capacities, that is, capacities for appreciating and living in light of what actually matters. On this way of filling in the story, what is crucial about persons is that they can do various things like form an evaluative conception, adopt and live in light of principles, engage in critical reflection and self-audit, and so on. The second puts the ability to be in touch with what really matters at the center. On this way of filling in the story, what is crucial about persons is that they have capacities for discerning and conforming to substantive normative truths about values and reasons. My goal in this dissertation is to plug for this second way of filling in the story and to explore the rich possibilities of building an account of autonomy around the idea that the agency of which persons are capable is best characterized in terms of substantively rational capacities.

Conclusion

This chapter began by introducing the challenge of conceptual diversity and went on to explore a variety of interpretive possibilities. It introduced four concepts integral to our thinking about autonomy: self-governance, authenticity, independence, and responsibility-entailing freedom. The chapter then considered three rival interpretations of the conceptual space and argued for the plausibility of a research program associated with structured pluralism. Instead of trying to locate unity at the level of concepts or giving up on unity altogether, I argued that a promising alternative is to search for theoretical integration. If an account can deliver theoretical integration,

and do so better than rivals, that speaks in its favor. I will return to the challenge of integration in chapter 3 where I spell out the core ingredients of a substantive view of personal autonomy and demonstrate its potential as a unifying theoretical framework. Before turning to that task, we need to consider the rival family of views which attempt to do without substantively rational capacities. That will be the challenge of the next chapter.

Chapter 2

The Trouble with Formal Views of Autonomy

There is a deep theoretical rift between formal and normative capacity accounts of personal autonomy. According to formal accounts, personal autonomy consists in conditions which can be specified in purely structural or procedural terms.[15] According to normative capacity accounts, personal autonomy consists, at least in part, in the possession of capacities for recognizing and responding to the norms that apply to one's choices and attitudes.[16] The first type of view denies that there are any substantive constraints on autonomously formed preferences and attitudes; the second affirms such constraints. In deciding on an account of personal autonomy, the choice between formal and normative capacity accounts represents an important fork in the road: it is one the deepest and most consequential choice points for our understanding of the nature of autonomy.

Formal accounts of personal autonomy represent the dominant type of view in the existing literature.[17] It is not difficult to see why formal accounts have seemed attractive to many philosophers: they seem to avoid controversial assumptions about normativity and metaphysics and steer clear of undesirable political implications like perfectionism and paternalism. Notwithstanding their attractions, formal views have troubles of their own—troubles which are rarely noticed. As I will explain below, such views have difficulty making sense of the idea that autonomy entails a fairly robust form of responsibility, seem committed to an arbitrary asymmetry

[15] Christman (1991a, 1991b, 2005, 2009), Dworkin, G. (1988), Ekstrom (2005), Frankfurt (1971, 1999), Friedman (2003), Killmister (2018), Meyers (2004, 2005), Westlund (2009).
[16] Benson (1987, 1990), Kauppinen (2011), McDowell (2010), Sayre-McCord & Smith (2014), Sher (1997), Stoljar (2000).
[17] Cf. Mackenzie and Stoljar (2000: 13), Westlund (2009: 26).

between the relevance of facts and values, and cannot properly vindicate the thought that autonomy is reason-giving in roughly the way we take it to be.

This chapter makes a case for reconsidering mainstream views of personal autonomy. It highlights several problems with formal accounts while arguing that the normative capacity alternative need not have the politically troubling implications it is sometimes thought to have. The last chapter ended by suggesting we search for theoretical integration in a story about normative agency. There are, I suggested, two broad possibilities for how that story might go. These possibilities correspond to the two families of views that are the focus of this chapter. Consequently, if this chapter succeeds in showing that formal views face serious challenges, it should help motivate us to search for a way of filling out the story that incorporates normative capacities.

The chapter has four sections. Section 1 begins with a brief characterization of formal views. Sections 3 through 5 articulate several *prima facie* objections to these views: that they cannot furnish an adequate account of responsibility, that they introduce *ad hoc* asymmetries between the importance of facts and values for autonomous agency, and that they are ill-equipped to vindicate the normative role played by the idea of autonomy. Section 6 explains why normative capacity accounts need not be inconsistent with liberal commitments.

1. Formal Views

Formal views of personal autonomy come in different shapes. The most popular variants offer some twist on the idea that to be autonomous one must be in some way identified with one's preferences and attitudes, for example, through taking reflective ownership of them or satisfying

the condition that one wouldn't disavow them if one became aware of their source (Christman 1991, Dworkin, G. 1988, Frankfurt 1971, Friedman 2003). Some of these views also incorporate external conditions which must be met, e.g., that the formation of preferences occur in the absence of coercion, manipulation, domination, and so on. Gerald Dworkin's (1988) classic statement of a formal view combines these two elements. On Dworkin's view, personal autonomy consists in "[…] a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values" (20). Choices must issue from this capacity, but as Dworkin makes clear, they must also do so in conditions of "procedural independence" (16, 18, 20), that is, in the absence of autonomy-undermining conditions like coercion and manipulation.

The defining feature of formal views is that they exclude substantive elements by design. They do so at two levels (cf. Benson 2005). First, they place no *direct* constraints on the contents of choice: *any* choice can in principle be autonomously made. Second, they exclude *indirect* constraints on choice in the form of substantively-defined attitudes or capacities featuring in the background. This means, for example, that on formal views autonomy cannot require that substantively-defined attitudes, like self-esteem or self-respect, feature in the background of choice, as some philosophers have proposed (Benson 2005). It also means that they cannot make substantively-defined capacities, like the ability to appreciate and respond to genuine values and reasons, a condition of autonomous choice, as normative capacity accounts maintain.

To be sure, on many formal views, autonomy does require some kind of rational capacity. For example, it may require at least thin, procedurally-defined, rational capacities like the ability to be sensitive to coherence constraints on beliefs and desires (Christman 1991a). According to another popular suggestion, autonomy requires agents to be able to treat considerations *as* reasons.

This thought can then be cashed out in functional terms. On a psychological version of the suggestion, considerations are treated as reasons when they play a specified role in the agent's psychic economy (Bratman 2009). On a social version, considerations are treated as reasons when agents are prepared to answer for themselves in the interpersonal exchange of reasons (Westlund 2009). Crucially, however, on formal views there is no requirement that agents be, or have the capacity to be, attuned to *genuine* reasons. Indeed, there *cannot* be such a requirement consistent with the strictures of formalism. Since they are committed to doing without substantive commitments in specifying the criteria of autonomous agency, formal views are unhitched from objective values and reasons by design. Fidelity to the core commitments of formal accounts therefore requires that whatever rational facility is criterial of autonomous agency, it cannot be reasons-responsiveness in any sense that requires being hooked up to genuine and substantive normative features independent of the agent.

A caveat about this definition is in order. Formal views, I have suggested, rule out direct *and* indirect substantive constraints on choice. In some of the literature on personal autonomy, however, the focal contrast concerns only the first level: it is between views that are *directly* substantive and those that are not. For example, Dworkin (1988: 12) contrasts his own view with a substantive account of autonomy, which he understands as placing *direct* constraints on what can be autonomously chosen. Similarly, Friedman (2003: 19) characterizes her view as "neutral with regard to the content of what a person must choose in order to be autonomous," and contrasts this with a substantive view according to which "someone is not autonomous unless she chooses in accord with certain values." The contrast invoked by Dworkin and Friedman does not perfectly align with the more demanding, two-level definition I have given.

Defining formal views in the more ambitious way is nevertheless appropriate for several reasons. First, it represents a trend internal to theorizing about autonomy by formal theorists themselves (Christman 1991b, Westlund 2009). This trend is in the spirit of non-substantive accounts of autonomy like those developed by Dworkin and Friedman, making explicit what these authors left implicit, or at any rate articulating a more thoroughgoing version of formalism. Second, the more ambitious two-level definition of formal views I have given follows more recent efforts at taxonomizing. The binary contrast invoked by Dworkin and Friedman does not adequately capture the space of interesting possibilities. For example, Paul Benson (1987, 1990) and Susan Wolf (1990) give accounts of autonomy on which it partly consists in the possession of reasons-responsive capacities. Benson (1994, 2005) has subsequently abandoned his earlier view and now defends the idea that autonomy requires certain substantively-defined attitudes, like self-respect or a sense of self-worth. Both types of view have a claim to being substantive in an interesting sense, even if they do not require that an agent make specific kinds of choices. For good reason, therefore, both types of view are now routinely classified as substantive (cf. Benson 2005, Christman 2018, Mackenzie & Stoljar 2000, Stoljar 2018). This means formal views are best seen as those that exclude substance at the second and not merely at the first level, i.e., at the level of capacities and attitudes as well as the content of choice. Third, by focusing on a more thoroughgoing formalism, the two-level definition presents us with a sharpened and more interesting contrast. Few accounts in the literature are directly substantive in the way envisioned by Dworkin and Friedman—and for good reason. So far as I can see, such accounts do not seem highly compelling. By contrast, indirect substantive accounts of the kind proposed by Benson and Wolf have a good deal going for them. The binary contrast invoked by Dworkin and Friedman risks obscuring the most interesting and relevant alternatives.

Before moving on, it is worth clarifying that the critical upshot of some of the arguments presented below has relevance for views which are not strictly formal. My criticism targets views which exclude normative capacity (or reasons-responsiveness) as a condition on autonomous choice. Consequently, insofar as the problems identified below are genuine, they will affect all views which exclude (or do not include) such capacities. Hence, views like Benson's (2005), which require attitudes like self-trust or self-respect but do not require substantive normative capacities, are vulnerable to many of the criticisms identified below. I nevertheless focus on formal views because these are popular and represent the starkest, and most thoroughgoing, alternative to thinking of personal autonomy in terms of the possession of normative capacities. They therefore constitute the most natural paradigm with which normative capacity accounts can be contrasted.

With these clarifications in place, let's turn to some problems with formal views of autonomy.

2. The Responsibility Challenge

A central problem with formal views of autonomy is that they cannot deliver an adequate conception of responsibility. In this section, I argue that there is a strong conceptual link between autonomy and responsibility and that reflecting on how best to interpret the notion of responsibility speaks against purely formal views and in favor of normative capacity accounts.

Begin with the link to responsibility. The association between autonomy and responsibility is widespread in the literature.[18] This is no accident; it reflects important conceptual connections

[18] For a small sampling: Arneson (1980: 475), Buss (2012: 648), Dworkin, G. (1988: 20), Dworkin, R. (1999: 224), Friedman (2003: 21-22), Westlund (2009: 30-36).

between the two ideas. Whether it is made explicit or not, a basic assumption in much theorizing about autonomy is that autonomy is responsibility-entailing in roughly the following ways:

(i) An autonomous agent is a responsible agent.

(ii) An autonomous agent is responsible for her choices and actions insofar as they issue from relevant autonomy-supporting capacities and circumstances.

(iii) All else equal, the greater one's autonomy in respect of choices and actions, the more one is responsible for those choices and actions.

Why think autonomy is responsibility-entailing in the sense expressed by these three claims? Let me highlight several pieces of evidence in support of this conclusion.

Performance Respect. The exercise of autonomy capacities typically merits a kind of appraisal respect related to the quality of an agent's performance.[19] When an autonomous agent enjoys circumstances conducive to the exercise of her autonomy-capacities, she merits our esteem or disesteem on the basis of how she exercises her autonomy. I use the language of esteem and disesteem here because it has fewer, or at any rate less narrowly, moral connotations than praise and blame and leaves open the precise connection to moral praise and blame. (On some views, what I am here calling esteem and disesteem will turn out to be a kind of moral praise and blame). What I have in mind is a credit-implying reactive attitude which tracks the exercise of capacities. While there are forms of appraisal that don't assume any notion of responsibility (e.g. appraising someone's physical attractiveness), other forms do, and it is quite plausible to think that exercises of autonomy are of this kind. At least in this context, esteem and disesteem are crediting responses, and they suggest that we see the agent as in some meaningful sense responsible for her choices

[19] See Darwall (1977) for the distinction between appraisal respect and recognition respect. Note that what I am here calling performance respect is only *part* of what Darwall (1977) calls appraisal respect. According to Darwall, appraisal respect includes assessments both of how agents perform in various roles/practices *and* of their characters.

and behavior. When we respond to a person with esteem or disesteem on the basis of how she exercises her autonomy, we plausibly see her as meriting such responses *via* her exercise of responsible agency.

The ethics of paternalism. It is plausible to think that at least part of what makes paternalism presumptively wrong is that it in some way violates autonomy (Christman 2018, Darwall 2006, Feinberg 1986, Groll 2012). In a suggestive metaphor due to Joel Feinberg (1986), autonomous agents enjoy a kind of self-sovereignty. Somewhat like the inappropriate meddling by one nation in the internal affairs of another, paternalistic interventions are thought of as illegitimate incursions into a person's proper sphere of choice. While one might interpret the sovereignty metaphor as suggesting that (hard) paternalism directed at competent adults can never be legitimate—that sovereignty sets an absolute side-constraint—a weaker claim seems at least as plausible: competent adults are entitled to strong presumptions against paternalistic interference, making it difficult to justify warranted interferences for their own good. (Note that in the international arena, sovereignty is not plausibly absolute either.) The idea that competent agents are entitled to a sphere of choice is the idea of autonomy as a *right* (cf. Feinberg 1986). As I said above, how people exercise their autonomy capacities is associated with performance respect; by contrast, their right to make choices as they see fit, is associated with recognition respect. To treat persons as little sovereigns is to treat them with due regard for their status as the kind of agents that merit protections against paternalistic interference.

This story seems to me hard to get off the ground without a background assumption about responsibility. Ideas about responsibility are plausibly implicated both in the scope and ground of the presumptive claim against paternalism. First, it is responsible agents who merit special protection against paternalistic interference. This is presumably why we think it is presumptively

wrong to paternalize adults but not children. The difference is that adults are responsible agents whereas children are not. Second, facts about responsibility affect the case for and against paternalism, so that (all else equal) paternalism becomes harder to justify as responsibility increases and easier to justify as responsibility decreases. This is presumably part of the reason why it is much easier to justify soft paternalism. Think of Mill's classic example of a man about to walk over a bridge he doesn't know is unsafe. Paternalism is easier to justify in such a case than it is to justify in the case where the man, knowing the bridge is unsafe, intends to walk on it. This is so *whatever* one thinks about the all-things-considered justification of paternalistic intervention in the two cases. The case for paternalistic intervention is stronger in the first case than in the second, and a natural explanation for this is that facts about responsibility are salient: given his ignorance, the first man is less responsible for his choice (and the outcome of that choice) than the second.

If part of what makes paternalism presumptively wrong is that it violates autonomy, we have here a powerful reason to think autonomy is responsibility-entailing. To be sure, one could coherently accept that paternalism violates autonomy and that the case against paternalism is sensitive to facts about responsibility while denying any connection between the two. One might, for example think paternalism involves the double wrong of violating autonomy *and* being inappropriately sensitive to facts about responsible agency. But this needs motiving. Antecedently, the simpler explanation connects the two: paternalism violates autonomy and autonomy entails responsible agency; it is the fact that agents are capable of being responsible for their own lives and choices that (in part) makes them autonomous; it is this very same fact that grounds a strong claim to being left free to pursue their lives and choices as they see fit. This picture is elegant in its simplicity, and it forges a straightforward connection between autonomy and responsibility.

Two brief caveats about the picture are in order. First, autonomy can imply responsibility without entailing that responsibility is sufficient for autonomy. There might well be additional elements to autonomy and, therefore, additional wrong-making features to paternalism. It is common, for example, to distinguish autonomy as a right from autonomy as an agency ideal. Plausibly, autonomy as an agency ideal is more demanding than mere responsible agency. But it nevertheless entails responsible agency. For the ideal to be in the offing, one has to be a responsible agent, capable of being responsible for one's life choices in some suitably rich and meaningful sense. Second, a well-justified regime of anti-paternalist norms presumably has multiple sources of justification. My point is not that considerations of responsibility are the exclusive source of justification for anti-paternalist norms, but that they are one important plank in the ethics of paternalism. Moreover, it is only fair to acknowledge that justifications of anti-paternalist norms are conceivable which make no appeal whatsoever to responsible agency (e.g., that agents typically know best what is in their interest). To fully defend the claim that strong anti-paternalist norms are best justified by a background assumption of responsible agency would require showing that alternative justifications, which do without the assumption of responsible agency, are not sufficient. That is more than I can do here. I'll therefore content myself with making the bet that these alternative explanations fail. They may contribute to partial justifications for anti-paternalist norms, but it is doubtful they can deliver complete and adequate justifications.

When someone decides to smoke or climb dangerous mountains, that choice plausibly merits respect *as* the choice of a responsible agent. Hence, as noted above, when someone is adequately informed about the risks, there is also stronger reason to desist from interfering. This is a backward-looking responsibility rationale: all else equal, there is more reason to allow persons to reap the consequences of their actions when they are undertaken responsibly than when they are

not. One might prefer a more forward-looking responsibility rationale instead (cf. Vargas 2013). Perhaps a regime of anti-paternalist norms can be partly justified by its proleptic or educative effects, tending to cultivate capacities for responsibility of roughly the kind it appears to assume. Either way, without the idea that persons are, or can become, responsible choosers, it is very difficult, I think, to support quite robust and general anti-paternalist presumptions of the sort most people in liberal societies subscribe to. The point here is not that people always live up to this picture of responsible agency or that facts about responsibility entirely settle issues about the ethics of paternalism. Rather, it is that our commitments to anti-paternalist norms plausibly depend on a deep background assumption of responsible agency.

Options. The third line of evidence comes from the persistent attractiveness of the idea that options matter for personal autonomy.[20] Raz (1986) gives memorable examples. The man who falls into a pit and can only decide when to nap or which direction to move his head is not very autonomous. Nor is the woman who is trapped on an island with a hungry beast and who spends her every waking moment trying to avoid being eaten by it. Something similar goes for the slave, who lacks options and cannot choose his own course through life (Oshana 1998), as well as for the many more prosaic forms of impoverishment which may not involve domination but nevertheless involve restricted options (Nussbaum 2001, Sen 1999). For example, it seems natural to describe refugees trapped in refugee camps as suffering diminishment of autonomy (Betts and Collier 2017).

Lack of options is constraining; it leaves agents less free to choose their course. It thereby also tends to diminish responsibility. Those who lack adequate options will tend to be less responsible for their choices and for the consequent shape of their lives (cf. Hurka 1987). The man

[20] Hurka (1987), Kauppinen (2011: 284ff.), Mackenzie (2014: 28), Mackenzie and Stoljar (2000: 22, 26), Oshana (1998: 94; 2006), Raz (1986), Terlazzo (2016).

in the pit is responsible for a few things—for whether he naps now or later, for which way he turns his head. But he is not responsible for much else about his life. His constrained circumstances change how it is appropriate to appraise the man. Before he fell into the pit, it might have been appropriate to feel some disesteem for him because, while enjoying significant talent and opportunities, he spent most of his days playing video games. Now that he is trapped in the pit, however, it would be absurd to feel disesteem for him on the basis of his unambitious choices. Because he lacks opportunities to exercise his agency capacities in a meaningful way, such performance-based assessments would be out of place. The impact of limited opportunity on *moral* accountability is familiar from fair-opportunity accounts of moral responsibility (cf. Brink and Nelkin 2013). Something similar seems plausible in the case of personal autonomy. In general, lack of options will tend to spell diminishment of autonomy. If personal autonomy implies responsibility, we can make sense of this. Limited opportunity undermines or threatens autonomy because, all else equal, it makes persons less responsible for their choices and lives.

Self-authorship/self-creation. A final piece of evidence for the link between autonomy and responsibility is to be found in widespread appeals to tropes of self-authorship and self-creation throughout the autonomy literature.[21] These metaphors express something deep and important about what it means to be autonomous, yet they are hardly intelligible without some background idea that persons are responsible for their lives. Creators and authors, after all, must be more than merely *causally* responsible for the products they create or author. To be self-authors or self-creators in any meaningful sense, persons must enjoy the right kind of responsibility-conferring relationship to their choices and lives.

[21] E.g., Benn (1976: 125, 127), Dworkin, R. (1993: 224), Enoch (2017: 27), Griffin (2008: 150), Raz (1986: 369-370, 390).

Together, these four lines of evidence suggest significant connections between personal autonomy and responsibility. In particular, they suggest that it is plausible to think of autonomy as responsibility-entailing in roughly the way suggested: that to be an autonomous agent, one must be a responsible agent; that an autonomous agent is responsible for her choices and actions when they issue from favorable circumstances; and that, all else equal, greater autonomy in respect of choices and actions implies greater responsibility for them. If the connections we have noticed are genuine, it is little wonder that the idea of responsibility crops up with some frequency in discussions of autonomy. There are systematic pressures supporting the idea that autonomy is responsibility-entailing.

Now for the trouble. Many formal theorists accept that autonomy comes with responsibility. Consider a representative quote from Gerald Dworkin (1988: 20): "By exercising [their capacities for autonomy], persons define their nature, give meaning and coherence to their lives, and take responsibility for the kind of person they are." The question is whether they have the resources to make sense of this commitment. More specifically, the question is whether they can deliver a notion of responsibility that is adequate to the task. Formal views spell out the conditions of autonomy in terms of properties like structural mesh between attitudes of higher and lower orders, agential coherence, actual or counterfactual reflective endorsement, forming temporally extended plans, treating considerations as reasons in the evaluation and adoption of plans, and so on. Such properties do seem well-suited to furnishing the basis for *some* ascriptions of responsibility. In particular, they seem to support judgments of attributability, according to which agents are related to their actions in such a way that their actions manifest their character and commitments.[22] In the case of moral conduct, responsibility-as-attributability typically means

[22] Following Gary Watson (1996), many discussions of moral responsibility distinguish two senses of responsibility: attributability and accountability. Roughly, one is responsible in the attributability sense if one's actions reflect one's quality

that actions express an agent's quality-of-will. But the idea of attributability can be generalized to cases not limited to moral matters. An agent will be attributively responsible for her life choices (even purely self-regarding ones) if they reflect on her—on what kind of person she is, on her sense of self, on her character, priorities, commitments, and values. Many formal views of personal autonomy are preoccupied with authenticity conditions. These aim to tell us when some choice or attitude is the agent's own in a special sense. Such accounts therefore seem to be well-equipped to capture the sense in which people can be attributability-responsible. When agents meet the requisite authenticity conditions, they stand in the relation of ownership to their choices and attitudes such that those choices and attitudes reveal where the agent stands, what she is about, and so on. Such views can therefore yield an important sense of responsibility: the kind which reveals something of the agent's inner life, putting her on display and opening her up to certain forms of appraisal.

The crucial question is whether this conception of responsibility is the right kind. Is it adequate for an account of personal autonomy? Two considerations suggest it is not.

First, as we have seen, the exercise of autonomy capacities typically merits performance respect. Since formal views of autonomy can plausibly make sense of attributability-responsibility, they plausibly have the resources to make sense of certain forms of appraisal respect: character-grading, aretaic judgment, assessment of motive, revealing where the agent stands, and so on. But performance respect requires something more specific. When an agent merits our respect for the exercise of her autonomy, her actions must meet a kind of credit-condition such that the agent can

of will, and one is responsible in the accountability sense if one's actions meet whatever control-conditions are required for being held morally accountable. A plausible specification of the control-conditions involved in moral accountability is that they consist (at least in part) in the possession of normative capacities (Brink & Nelkin 2013, Fischer and Ravizza 1998, Nelkin 2011, Wolf 1990). By contrast, a plausible specification of attributability-relevant conditions requires only that an action reflect something like the agent's genuine or authentic self—her character, perspective, or will.

earn our esteem or disesteem on the basis of how she exercises her autonomy. Some crediting responses are quite weak: they amount only to something like approval or disapproval. Attributive responsibility suffices for making this weak class of responses apposite. Other crediting responses, however, are stronger: they amount to something like performance-criticism. It is not clear attributive responsibility suffices for this stronger class of responses.

Consider that the facts which determine choice-worthiness are normative. Since the paradigm of personal autonomy is often taken to be self-regarding choice, consider for simplicity the domain of prudence. On all of the most widely held and plausible views of welfare, there are facts about what is good for agents that is independent of their momentary desire and whim. Choices in this domain can be better and worse, right and wrong, wise and unwise, and so on.[23] It is hard to see how the stronger class of crediting responses could be apposite in the absence of sensitivity to the very facts that determine choice-worthiness. To be a suitable target of performance-criticism on the basis of how an agent exercises her autonomy capacities, she must enjoy the right kind of control. But it is hard to see how the agent could enjoy such control in the absence of normative capacities. Agents who satisfy formal autonomy criteria but lack normative capacities seem a bit like blindfolded dart throwers attempting to hit a target.[24] Why should an agent merit our disesteem if she is completely insensitive to the facts in virtue of which she ought to choose one way or the other or cannot suitably regulate her conduct in light of this sensitivity? Unhitch agents from the relevant normative facts, either because they are blind to them or incapable of acting on them, and it becomes very hard to see how they can be responsible for their

[23] This is true on hedonist, objective list, and perfectionist views, but it is true on the most compelling versions of the desire-satisfaction view as well, which add counterfactual and idealizing conditions as a filtering mechanism on which desires the fulfillment of which count toward a person's welfare.
[24] For the metaphor of blindness, see Kauppinen (2011: 281), Wolf (1990: 92).

choices in the way that is characteristic of the kind of performance respect we associate with the exercise of autonomy.

Second, as we also have seen, on a plausible interpretation of the ethics of paternalism, both the scope and grounding of anti-paternalist principles is sensitive to facts about responsible agency. Does being attributability-responsible suffice to ground robust anti-paternalist norms? It is hard to see how it could. The same facts that make it difficult to see how an agent who lacks normative capacities could have the kind of control needed to render performance-criticism apposite also make it hard to see how it could ground a strong claim against paternalistic interference: it is precisely because children lack such capacities that they do not have a strong claim against intervention by parents and educators.

Formal views of autonomy do frequently posit reflective ownership capacities. Would such capacities suffice to merit anti-paternalist protections? It is hard to see how. Again, the domain of choice is governed by practical norms. Stipulate that the agent is insensitive to these norms and it becomes difficult to see why she merits strong protection against intervention by third parties. To be sure, since facts about responsibility are not the only relevant facts for determining the appropriateness of paternalistic intervention, there may still be all-things-considered reasons to protect her choice even if she is not sufficiently responsible. But, as I argued above, facts about responsible agency are a huge pillar in the anti-paternalist case. Once this is acknowledged, we need an interpretation of the relevant notion of responsibility. What we need is a kind of responsibility that is robust enough to ground strong anti-paternalist norms and (as a corollary) puts agents on the hook for the upshots of their own choices. Mere identification with, or ownership of, attitudes seems much too weak. What matters is not that a choice is authentically yours; what matters is that the choice is one for which you are robustly responsible, in such a way that you can

be on the hook for its consequences and that I have strong presumptive reasons to let you make your choice even if I could do better. To secure this result, something more is needed. Adding reflection does not do the trick. Perhaps reflective endorsement increases authenticity, such that you then own the choice in a special and deeper way. This may say something about you—about your character, perspective, and values. In that minimal sense, it constitutes a kind of responsibility. But reflective endorsement does not amount to enjoying a relation to your choices that would explain why other agents have presumptively decisive reasons to let you have your way, even when they know better. The fact that a choice is *authentically yours* just doesn't seem to have the right kind of normative relevance to ground such reasons. By contrast, if you have normative capacities for appreciating and responding to the values and reasons bearing on your choice, then that does seem to do the trick. If you can appreciate and respond to the normative features relevant to your choice, then that gives you a deeper kind of control over your choices and actions, putting you on the hook for their upshots, and giving me reasons to desist from paternalistically interfering with your choice.

To return to the difference between children and adults: what is the salient difference between children and adults, such that paternalism of the former is generally more acceptable and paternalism of the latter? It is plausibly a difference in their status as responsible agents. But what kind of responsibility is relevant here? I have argued that strong anti-paternalist protections would be better supported by a form of responsibility that puts agents on the hook for the upshots of their choices than a form of responsibility that merely reveals what kind of person they are, where they authentically stand, etc. If that is right, the relevant difference between children and adults seems to be that the former have more fragile normative capacities than the latter, not that they have a less well-developed sense of self. This has some intuitive plausibility. Think of it this way. You

are about to meet a 7-year-old child who is a stranger to you. All you know is that the child is extremely precocious and wants to undertake a dangerous activity. You have the power to stop her. Which set of facts would ground a stronger claim against you not to interfere with her choice? The fact that she is reflectively mature and seems to have crystalized a perspective and stance on the world that is genuinely her own? Or the fact that she is reflectively mature in such a way that she seems sensitive to normatively relevant features of her choice? I suspect you will agree that normative capacities ground stronger claims against intervention than mere authentic ownership of choices. While children generally have more fragile normative capacities and a less developed sense of self, it is the first of these properties that seems more important in considering whether paternalistic treatment is warranted.

Together, these considerations put enormous pressure on formal views of autonomy. What we need, I have argued, is an interpretation of autonomy that can deliver a robust conception of responsibility. The worry is that formal views cannot deliver such a conception. They can give us a conception of responsibility which shows us where the agent stands and thereby reveals something good or bad about her. But they cannot give us a conception of responsibility which shows the agent to be in the kind of relationship to her attitudes and choices that seems to be required by our treating exercises of her autonomy as meriting positive or negative performance respect; nor can they give us a conception of responsibility robust enough to ground strong claims in favor of allowing the agent to live with her choices and against others that they not paternalistically interfere.

3. The fact/value asymmetry

Perhaps formal theorists will succeed in giving us a rich and convincing story about responsibility. Even so, a further problem looms. Responsibility requires adequate non-evaluative information. This idea is familiar from discussions of *moral* responsibility, where ignorance is typically taken to be exculpatory. Ignorance, of course, does not *always* excuse, as in the case of willful or negligent ignorance, but ignorance can and frequently does serve as an excusing condition in assessments of moral responsibility. The analogous thought is plausible in the case of personal autonomy as well: just as inadequate information can diminish moral responsibility, inadequate information can diminish autonomy. Someone who smokes in complete ignorance of the risk this poses to her health is plausibly less autonomous with respect to that choice than someone who is apprised of the facts and chooses to smoke anyway; a lover who marries her beloved ignorant of his true character is less autonomous with respect to that choice than someone who knows her lover in greater depth; and so on. All else equal, more choice-relevant information means more autonomy; less choice-relevant information means less autonomy.

It is possible to deny that non-evaluative information is relevant to autonomy. Michael McKenna (2005) commits himself to this bold thesis in an effort to discover some interesting difference between moral responsibility and personal autonomy. In his central example, Tal attempts to help his sick friend, Daphne. Pulling mislabeled medicine from the cabinet, Tal gives Daphne poison and thereby accidentally poisons her. According to McKenna, Tal acts autonomously (though he is not morally responsible) in poisoning his friend. This is because, explains McKenna, there is a sense in which Tal rules himself by acting in accordance with self-chosen principles. The principle on which Tal acts is: always attempt to help those who suffer innocently. And Tal's action conforms to this self-chosen principle because, in administering the

drug, he does attempt to save his friend. The resulting picture is a fairly stark form of internalism on which autonomy is compatible with sweeping ignorance of relevant facts.

This is not compelling. On its own, the example seems to lend intuitive support to the thought that Tal's autonomy is undermined or threatened by his ignorance. So do similar examples discussed by Al Mele (1995: 179-181), like the example of Connie, who chooses an investment plan but is systematically deceived by the company offering the plans, and King George, who rules his kingdom contrary to his deepest commitments because his staff systematically distorts the information arriving at his desk. If anything, it seems intuition antecedently favors the verdict that these agents suffer some impairment of autonomy by being informationally cut off. McKenna acknowledges the intuitive pull of Mele's examples, but he insists that the intuitive pull tracks moral responsibility rather than personal autonomy. If we stipulate that autonomy is acting in light of self-chosen principles, then, suggests McKenna, Tal and Connie and King George can all be seen as autonomous. But this is not obvious. Even granting McKenna's stipulative definition of autonomy, no strong form of internalism follows, for it is plausible to suppose that acting in light of one's principles imposes success conditions on action which are not met in the examples.[25] McKenna avoids this problem by describing Tal's action (and by implication, Connie's and King George's) as an *attempt*. Tal's principle is: *attempt* to help your friends. This is something he succeeds in doing. But suppose his principle were: help your friends. This is not something he succeeds in doing. The intuitive force of the examples as instances of autonomy thus depends on an artefact of description. Tal and Connie and King George would, of course, not be identified with their actions under an informationally enriched perspective. Once we shift the act description to a more objective frame, it becomes much less compelling to think of them as autonomous.

[25] For similar points of criticism, see Killmister (2013).

Consider another possible fix. One might describe all principles of action in evidence-relative terms. Tal's principle might be: act to fulfil your goals and values as indicated at the moment of action by your subjective evidence base. He might then, for example, grievously harm Daphne while counting as exemplary in his autonomy simply because he acts on his (misleading) evidence about what helps and harms her. But what would serve the values and principles of agents is not typically evidence-relative or subjective in this way: Tal cares about his friend, Connie cares about her future in retirement, King George cares about the flourishing of his kingdom, and so on. This outward focus imposes objective success-conditions which require being suitably well-informed if one is to promote the relevant values and principles. Antecedently, as I said, intuition seems to favor Mele's verdict about the cases, that autonomy is threatened by deprivation of decision-relevant information. One can try to deflate some of the intuitive force of these examples, as McKenna seeks to, but only by re-describing the principles and values from which agents act in terms of implausibly unambitious success-conditions—as attempts or evidence-relative respondings. If one sticks with a realistic interpretation of what agents actually care about, then their being autonomous plausibly does depend on being adequately informed.

As we have seen, there are strong conceptual and theoretical pressures to preserve the association between personal autonomy and responsibility. McKenna arrives at his conclusion precisely in an effort to locate some interesting notion of personal autonomy that *comes apart* from responsibility. But he offers neither theoretical motivation nor robust conceptual anchor points for this strong internalist suggestion. If, as I argued above, autonomy is a form of personal freedom in virtue of which agents are responsible, then we have good reasons to reject the kind of extreme informational hermeticism on which an agent can be completely ignorant or deceived about factual information relevant to her choice.

Theoretical pressure is increased by noticing the connection between being informed and having control. Ignorance threatens an agent's control (cf. Mele 1995). The examples of Tal, Connie, and King George exemplify this. By being significantly ignorant, these agents have impoverished control over their actions. And control seems fairly clearly relevant to autonomy. Think of a case involving complete absence of executive control. Perhaps in the inner sanctum of my mind I endorse normative principles and aim to conform my actions to them. It seems utterly implausible to think I enjoy autonomy if I have no power whatsoever to conform my actions to my principles. But if lack of control threatens autonomy on the "active" side, why wouldn't it do so on the "receptive" side as well? After all, both executive capacities and representational capacities can be thought of as aspects or dimensions of control. In the absence of reasons to posit an asymmetry in control conditions, it seems arbitrary and unmotivated to insist that one dimension of control matters while the other does not. In short, McKenna's proposal is an interesting suggestion about how to divide conceptual labor between moral responsibility and personal autonomy, but we have few independent reasons to accept it and some independent reasons to reject it.

Must we go to the other extreme and hold that only those who act in light of all relevant information are autonomous? This would entail the absurd conclusion that almost no one ever acts autonomously. We need not accept this extreme conclusion. An intermediary view is available, namely that *sufficient* information is necessary for autonomy. This is plausible if we distinguish scalar and threshold assessments of autonomy. Many of the properties relevant to autonomy (like being informed) are a matter of degree. Deploying a scalar conception of autonomy, we can say that agents are more autonomous the more they satisfy the relevant scalar property. Switching to a binary, threshold conception, we can say that some threshold level of the property must be

reached to qualify as autonomous. The two conceptions can be combined. On such a picture, autonomy "kicks in" above the threshold but one can be more or less autonomous (perhaps with no upper bound) above that point. When it comes to being informed, the combined conception seems plausible. Below some threshold of understanding, agents may not be autonomous with respect to a choice at all. Above that threshold, being more informed tends to enhance, and being less informed tends to diminish, autonomy. Citing Columbus's ill-informed decision to sail west, Arpaly (2005: 175) doubts "that anyone wishes to claim…that an ill-informed decision cannot be an instance of autonomous agency." But she also accepts that giving someone more information might make the person more autonomous. Once we distinguish scalar and threshold verdicts, both of these claims seem plausible. Being ill-informed can be autonomy-impairing while not rendering one entirely non-autonomous.[26]

Now for the challenge. Suppose formal views accommodate the idea that non-evaluative information is relevant to autonomy. By parity of reasoning, it seems plausible to suppose that responsibility likewise requires *normative* information. It is hard to see why someone who is completely normatively "blind" would be any more autonomous than someone who is ignorant of non-evaluative information. Suppose someone smokes, knowing the risk this poses to her health but in total ignorance of what is good for her or the reasons this gives her to make one choice rather than another. Such a person seems just as blind, in the relevant sense, as someone who is ignorant of the non-evaluative facts. There is, then, a simple parity argument for treating factual and normative ignorance alike.[27] If autonomy implicates responsibility, and if both factual and

[26] So far as I can see, the *source* of misinformation is irrelevant to autonomy. Tal's autonomy is not lessened more if his misinformation is the result of intentional manipulation than if it is the result of accidental labeling. Similarly, if Connie and King George are informationally impaired due to a fluke of circumstance, this is no less an impairment of their autonomy than if they are the victims of campaigns of disinformation.

[27] On the symmetry of facts and values in moral responsibility, see Rosen (2003) and Wolf (1990); for the parallel claim about personal autonomy, see Kauppinen (2011: 280) and Savulescu (1995: 330).

normative ignorance can defeat or attenuate responsibility, we have (in the absence of further considerations) no more reason to credit autonomy in the absence of the one than in the absence of the other.

Extreme internalist conceptions of autonomy are implausible. They suggest that complete factual ignorance does not in any way threaten autonomy. Formal theorists therefore do well to accept that non-evaluative information can make a difference to autonomy (cf. Berofsky 1995, Killmister 2013, Mele 1995). But once this much is accepted, there is pressure to accept that evaluative information is relevant to autonomy as well. If one accepts that autonomy is responsibility-entailing, there is a principled rationale for taking this further step. Sensitivity to evaluative information is just as relevant as purely factual information in constituting an agent as responsible. The domain of choice is one in which norms apply: choices can be better or worse, right or wrong, prudent or imprudent, and so on. Truths about choice-worthiness are a function, not of descriptive facts *per se*, but of descriptive facts *plus* relevant evaluative or normative truths. Hence, truths about choice-worthiness are partly normative. But truths about choice-worthiness also furnish the basis for critical assessments of agents. Factual and normative deficits alike tend to be responsibility-diminishing: below some minimal threshold-level, agents are not responsible for their choices at all; above that level, they are more or less responsible depending on their sensitivity to the relevant features.

Formal views, however, must reject parity. Since they sever the connection between autonomy and substantively defined evaluative capacities, such views must also deny that normative information matters for autonomy.[28] This creates an explanatory burden. On the face of it, the exclusion of evaluative information seems *ad hoc*. This puts pressure on formal accounts to

[28] Killmister (2013: 527) does just this, arguing that false factual beliefs tend to impair autonomy but not false principles or values.

explain why evaluative and non-evaluative information should be treated in an asymmetric fashion. The answer cannot be: because that is what is predicted by formal theories. In the absence of some salient difference we have independent reasons for accepting parity. This speaks against formal theories precisely because they predict an asymmetry. It is therefore not satisfactory to point to this implication of formal theories in reply to the challenge. Perhaps formal theorists can ultimately give us some principled, non-question-begging story about why we should treat normative and factual information differently. In the meantime, we have a *prima facie* case for thinking autonomy requires sensitivity to evaluative information. This speaks against formal views of autonomy.

4. Autonomy's Normative Role

The final worry about formal views is that they cannot make adequate sense of autonomy's normative role. We recognize autonomy's normative role in the kinds of reasons it supplies. Autonomy is reason-giving in roughly two ways. On the one hand, we think it good, all else equal, for people to live autonomous lives and make choices autonomously. It is therefore the sort of thing we have reason to aspire to ourselves and promote the realization of in others. On the other hand, we think autonomy marks out a sphere within which individuals are free to choose and that their autonomous choices carry authority or gravity in certain contexts of decision-making to which we must often give greater weight than the choices would merit on the basis of their direct consequential value or other forms choice-worthiness.[29] When an agent or her choices meet the

[29] One might object to putting the point in terms of the weight of reasons. According to Groll, an autonomous will is to be taken as "structurally decisive" (2012: 699-706). However, Groll suggests that paternalism is only "presumptively wrong" (710-711). So even on a Raz-style view like Groll's, where certain considerations are shielded from entering deliberation, the shield is not necessarily absolute.

conditions of autonomy, we must take her decision with special seriousness. Even when it is trumped by other considerations, autonomy places the bar of interference higher than it otherwise would be: it ratchets up the demands for warranted intervention.

An adequate conception of autonomy should be able to make sense of this twofold normative role. In other words, an adequate conception of autonomy needs to vindicate the thought that autonomy is worthy of promotion and worthy of respect. But there are reasons to doubt formal views provide ingredients sufficient to meet this demand. Let's consider each of these normative roles in turn.

What are the kinds of autonomy-relevant conditions we generally have reasons to promote? The most obvious is perhaps this: to ensure that people have sufficiently valuable options to choose from. Some autonomy theorist, like John Christman (2005: 282), deny that valuable options matter to autonomy.[30] But often the motivation for this denial is heavily theory-driven, for example, by the desire to avoid perfectionist implications. Pre-theoretically, it is quite natural to describe people with limited valuable options as suffering a diminishment of autonomy (cf. Nussbaum 2011, Oshana 1998, 2006, Raz 1986). Consider refugees living decades of their lives in a camp. These people typically have a dearth of valuable options and it is natural to think of them as suffering from an autonomy-relevant impairment as a consequence (Collier and Betts 2017, ch. 6).

Valuable options are an external good. We plausibly also have reasons to promote an internal good to go along with it. Think of what parents want for their children. Parents do not just want their children to face a lush banquet of valuable options; they want their children to possess the capacities to appreciate and respond appropriately to those options. This pattern of concern seems appropriate more generally. It would seem a bit odd to care that persons enjoy valuable

[30] Christman (2009: 170) subsequently argues that, if valuable options matter to autonomy this should be understood in terms of a subjective conception of value, i.e., the options need only be valuable from the perspective of the agent.

options but not to care that they enjoy capacities for appreciating and responding appropriately to those valuable options. Some formal theorists argue that autonomy requires valuable external options but that the internal capacities should nevertheless be understood in terms of purely procedural conditions (Terlazzo 2016). This is an unstable position. Once one accepts that valuable options matter, why not also accept that internal competencies for tracking and pursuing those valuable options also matter?

There are, of course, complicated questions about who may promote whose autonomy and how this may be done. Some liberal theorists, for example, insist that the state may play no role in promoting the autonomy of its citizens. Moreover, there are plenty of cases where we have reasons not to promote, and even to curtail, autonomy—for example, prospectively, when people's exercise of autonomy will likely bring about significant and unjustified harms to others, and retrospectively, for purposes of punishment. But the present point does not depend on denying such qualifications. What it depends on is only the broad generalization that people ordinarily have robust reasons to promote their own autonomy and often also the autonomy of others. The exceptions are important but they shouldn't obscure the fact that there are general standing reasons for anyone to promote anyone else's autonomy. A plausible interpretation of what people generally have reason to promote includes (i) valuable options, and (ii) normative competence over those options. It is consistent with this to think that there are secondary considerations excluding states or other agencies from the role of autonomy-promoter and that in some cases there is most reason *not* to promote autonomy.

This specification of what people often have reasons to promote fits elegantly with a normative capacity account. It does not fit well with formal accounts. When we think about the kinds of properties identified by formal accounts—reflective acceptance, the ability to treat a

consideration as a reason, answering for oneself in the social exchange—it is at least not obvious whether and why we have reasons to promote these things. Perhaps we do have reason to promote these things; much will depend on how the details are spelled out. But it surely isn't obvious that we have quite general and powerful reasons to promote these properties. Contrast this with our confident commitment to promoting autonomy. Barring complications about special secondary reasons some agents might have not to be autonomy promoters, we think there are standing agent-neutral reasons to promote *anyone's* autonomy. This confidence is readily vindicated if autonomy turns out to require (i) valuable options, and (ii) normative competence. We can readily appreciate why these twin goods would be valuable and worthy of promotion. Perhaps formal views can ultimately rise to the challenge of explaining why the properties they posit as constituents of autonomy are worthy of promotion. But the case needs to be made. There is at least a *prima facie* challenge here: normative capacity accounts are well-positioned to make sense of the idea that we generally have reasons to promote people's autonomy; formal accounts, by contrast, are not so obviously well-positioned—whether they can make sense of the reasons we have to promote autonomy is more of an open question.

Perhaps, however, this is an unfair assessment of the situation. Consider the following problem. There is an ambiguity in the idea of reasons-responsiveness: does it mean merely *having capacities* for responding to reasons or *actually exercising* those capacities? Which of these does the normative capacity account of autonomy appeal to? Is the mere capacity for responding to reasons enough for autonomy or must one also exercise one's capacities in such a way as to conform to one's reasons? The label—normative *capacity* account—certainly suggests the former. There are also systematic pressures encouraging this interpretation. Intuitively, it seems that choosing autonomously is not the same thing as choosing rightly or wisely or well. And this

intuition is underwritten by one of autonomy's central normative roles: if part of what makes paternalism presumptively wrong is that it conflicts with autonomy, it is hard to deny the possibility of autonomous bad choices; and if autonomous bad choices are possible, autonomy cannot consist in appropriately exercising one's capacities for reasons-responsiveness. Moreover, the successful-exercise-of-normative-capacity interpretation of autonomy seems to imply that only the virtuous are really autonomous. If one wants a view of autonomy that squares with standard liberal commitments, the pure capacity interpretation looks far more promising. But suppose the pure capacity interpretation of autonomy is right. Then it is no longer so clear why autonomy is the sort of thing we generally have reasons to promote.[31] Merely having *capacities* for reasons-responsiveness, after all, is not all that valuable; what *is* valuable is having the capacities *and exercising them well*, i.e., actually succeeding in responding to one's reasons. The normative capacity theorist cannot have her cake and eat it too: if she wants an account that makes sense of *one* of autonomy's key normative roles—being a bar to paternalism—she has to give up on being able to account for its *other* normative role—being the sort of thing we have reasons to promote.

Here are two possible responses. The first is to deny the objection's presupposition and to insist that mere normative capacity *is* valuable and worth promoting after all. How so? Put briefly, normative capacity constitutes persons as responsible in a deep and meaningful sense for their lives—and *that* is good. To be sure, being responsible need not always and invariably be good: maybe being responsible for very bad decisions can make someone's life go worse. However, this qualification is consistent with the general and prospective value of autonomy as a thing worthy of promotion. In chapter 4, I'll argue that normative capacities are valuable because a variety of important human goods are conditioned or amplified by being freely and responsibly pursued, and

[31] Thanks to David Brink for pressing me on this point.

I'll suggest that, while they can be misused and wasted, normative capacities are therefore significant *opportunity goods*. This vindicates, at least in a general way, the claim that we have reasons to promote normative capacity.

The second response accepts the objection's presupposition. It agrees that merely having normative capacities is not valuable and, hence, not the sort of thing we generally have reasons to promote, but it goes on to insist that this only shows that we should reject a purely capacitarian account of autonomy. For the reasons I have given, this response may look unpromising. But it need not be. The response would be unpromising if it simply collapsed autonomy into a form of virtue without remainder. It need not, however, do that. T.H. Green (1886/1986) distinguishes between two kinds of freedom: responsibility-entailing freedom and perfection-entailing freedom (Cf. Brink 2003: 81). Putting this in terms of normative capacity, the former idea is the idea of having the ability to detect and pursue norms and values—reasons-responsiveness. This ability is plausibly at the basis of responsible agency, so the corresponding idea of freedom is a responsibility concept: one is responsible and can, in whatever way is suitable, be held responsible for one's choices. This ability can be had even if it lies dormant or is exercised poorly. The latter idea is the idea of realizing normative capacity—of actually successfully tracking relevant norms and values and then successfully *conforming* behavior accordingly. This ability is plausibly a kind of perfection of our rational natures, so the corresponding idea of freedom is a virtue concept: one realizes an important human excellence and merits approbation and esteem on that basis. The second response can be put as follows: once we distinguish the responsibility concept from the virtue concept, we can see that what it is we have reasons to promote is the property corresponding to the virtue concept—perfection-entailing freedom. As long as we also accept that respecting responsibility-entailing freedom is important, we haven't collapsed autonomy into virtue; instead,

we have come to see that our thinking about autonomy is more complex than we might have initially thought.

It seems to me that the first response is partly right: there is *some* positive value to normative capacity as a generic opportunity good and this accounts for some of the reasons we have to promote autonomy. But it seems to me the second is partly right too: by itself, unexercised normative capacity is *not very* valuable; we also, and perhaps especially, have reasons to promote the fulfillment or appropriate exercise of normative capacity.

The ambiguity between *mere capacity* and *fulfillment of capacity* is present in many of the normative capacity accounts that have been offered in the literature. Once the ambiguity is noticed and the alternatives clarified, there is an important question about how best to develop such a view: should we go for a pure capacity view or a pure virtue view? The answer, I believe, is that we should reject the choice as a false alternative and give up instead on the assumption that autonomy is a unitary thing. Making use of the distinction introduced by Darwall (1977) between two forms of respect, we can recognize distinct normative statuses associated with our thinking about autonomy: recognition respect goes with being a responsible chooser, appraisal respect with how capacities are exercised. Moreover, once we spell out autonomy's normative roles, we see that it commits us to both the responsibility concept and the virtue concept, each one associated with a different status. As we have just seen, the idea of perfection-entailing freedom is needed to account for the full value of autonomy, that is, for all the reasons we have to promote autonomy. In the next chapter, I will offer a further argument in support of the conclusion that the virtue concept cannot be dispensed with, viz. that it tracks success with respect to an aim that is internal to autonomous agency. If these arguments are right, we need to accept a further element as part of our thinking about autonomy. This element was not part of the initial mapping of conceptual space

given in chapter 1, but on closer inspection we can see that normative pressures commit us to including it as part of our thinking about autonomy.

Turn now to reasons of respect. Respecting a person's autonomy means at least two things. First, it means respecting the person's right to make self-regarding choices as she sees fit, including (perhaps up to some threshold) bad choices. This idea is, of course, closely associated with anti-paternalist norms in liberal societies. Second, it means honoring the person's perspective—her wishes, what matters to her, what she cares about, and so on. The latter shows up, for example, in what is required to treat someone of another religion with respect or (a bit more broadly) in respecting claims of conscience, whether religious or secular (Maclure & Taylor 2011).

As we have seen, formal views of autonomy are commonly taken to be in the business of specifying conditions for authenticity. Suppose they succeed at this. Then it seems they have the ingredients for vindicating the second manifestation of respect, i.e., the one having to do with respect for conscience. Say you must decide whether to give a blood transfusion to an unconscious Jehovah's Witness to keep her alive. You know for a fact that she would not want to be given the blood transfusion, even though her life depends on it. It is not obvious what you ought, all things considered, to do. Still, whatever the right thing to do is, it seems there are powerful reasons of respect speaking in favor of honoring her (counterfactual) wishes not to receive the transfusion. Contrast this with a case where you know the religious commitments are superficial or have been inculcated in a suspect way. Perhaps the person has only been flirting with the Jehovah's Witness community for a couple of weeks or she has been drugged, manipulated, brainwashed, or coerced, into having the commitments she does. In this case, presumably the weight that should be given to respecting the person's wishes is less, if any should be given to them at all. The difference between these cases is not in the content of the patient's request—that's the same. Instead, it is to be found

in something like the position of the request vis-à-vis the person's authentic self (cf. Enoch 2017). Insofar as formal views are equipped to give us a story about authenticity, then, they are in a position to give us a story about this crucial dimension of respect for persons: honoring (i.e., giving some weight to) their point of view.

But it is not clear formal views have the ingredients for vindicating the first manifestation of respect, i.e., the one having to do with the strong anti-paternalist presumption. As I argued above, authenticity plausibly suffices for attributability-responsibility, but this is not the right kind of responsibility to make sense of strong anti-paternalist practices. The strong liberal anti-paternalist presumption would seem best justified by the assumption that persons are more than merely attributability-responsible for their choices. There are often good reasons *in favor* of paternalism, even when paternalism is all-things-considered wrong. In particular, people's welfare matters greatly and any balanced assessment of the ethics of paternalism must recognize this side of the balance sheet. How could a person's foolish choices merit protection? How could the kind of freedom that would allow people to make potentially ruinous life-choices be justified? The mere fact that a choice is authentically an agent's own would not seem to suffice to give other persons presumptively decisive reasons to desist from paternalistically interfering with the choice. A more robust form of responsibility seems to be required to make sense of that. This more robust form of responsibility would be secured by the ability to appreciate and appropriately respond to normative features relevant choice.

To be clear, my point is about the *presuppositions* behind our *general* stance. I am not suggesting that considerations about normative capacity always feature, or always ought to feature, or that they are the only or always the most important considerations, in every particular case in which the anti-paternalist presumption holds up. Nor am I claiming that there isn't *some* anti-

paternalist support from mere attributability-responsibility: there are, as I argued in the last paragraph, *pro tanto* reasons to honor people's points of view in self-regarding matters and these plausibly contribute to the case against paternalism. Rather, I am claiming that the kind of robust anti-paternalist norms characteristic of liberal social morality would be difficult to justify unless people were responsible (in the sense of being on the hook) for their choices. And this is difficult to make sense of in the absence of relevant normative capacities. Normative capacity is the ability to register and appropriately respond to normative features relevant to choice. Thus, if normative capacity is not required for autonomous choice, as formal accounts must maintain, this means one's choice about X can be autonomous independent of any sensitivity to the features in virtue of which X is choice-worthy. But this is surely puzzling. For how can one be robustly and meaningfully responsible for choosing X in the absence of capacities for tracking what is relevant to the question whether one should choose X? The requirement that we honor people's perspective is not strong enough for quite general and robust anti-paternalist norms. Perhaps such considerations weigh in here and there in particular cases, but they cannot plausibly be thought to support a regime of vigorous anti-paternalist protections. Assuming that paternalism is (at least in part) presumptively wrong because it conflicts with people's sphere of "sovereignty" (Feinberg 1986), and assuming that autonomy-as-a-capacity is the ground of autonomy-as-a-right, we need to ask what view of autonomy capacities would be required to justify a robust sphere of self-sovereignty. Views of autonomy which include a normative capacity condition would seem much better equipped than views which do not, to vindicate a robust sphere of self-sovereignty.

It is possible, of course, that people's *actual* normative capacities are often quite fragile. In that case, the idea that persons are normative agents, capable of tracking and responding to the normative features bearing on their choice may be something of an idealization. One might

maintain, as I suggested earlier, that the anti-paternalist norms are partly proleptic or educative, cultivating the thing they appear to presuppose. Or one might maintain that, although it is something of an idealization, people are responsible often enough and, given the other contributing reasons against paternalism and perhaps secondary reasons against too closely tracking facts about normative competence, it is an acceptable idealization. My point is not to defend anti-paternalist norms but to clarify what we are plausibly committed to in accepting them. If someone thinks adults are *not* normatively competent *most* of the time, then it seems to me they should in principle be prepared to accept a much more invasive regime of paternalism than we tend to think appropriate in liberal social orders (even if in practice such a regime would be too difficult or too expensive or too unpopular or too abusive, etc). For if adults are really not normatively competent most of the time, they will in this respect be a lot like children, and it will be difficult to see what principled objection would remain to treating them like children, except that there might be a variety of practical obstacles to doing so. If one thinks there are strong principled objections against paternalism—that the objections to it are not just incidental or technical—there is substantial pressure to also acknowledge that persons are, or can, be responsible agents. Even if our self-conception as responsible agents is slightly idealized, as long as it does not radically betray the facts about us, we can make sense of strong principled objections to paternalism. This self-conception of responsible agency—perhaps a mix of fact and ideal—is better captured by normative capacity accounts than by formal accounts of personal autonomy.

In sum, autonomy is a recognizable value in liberal social orders, which prize self-direction and which are committed to protecting a significant sphere for individual choice. At the heart of this social vision is the idea of persons as dignified choosers who must chart their own course through life (Dworkin 1988, Raz 1986). This idea marks out two normative roles for the idea of

autonomy. One is an agency ideal. All else equal, autonomy is a desirable agency characteristic. Another is a principle protecting the exercise of autonomous agency. The choices of an autonomous agent call for respect. These are familiar ideas. And my argument in this section has been that, on the face of it, the normative role played by our concept of autonomy fits much more naturally with normative capacity accounts of personal autonomy than formal accounts.

5. Are Normative Capacity Views Compatible with Liberalism?

Tension with perceived liberal commitments is a central source of resistance to normative capacity accounts. For example, John Christman (1988, 1991, 2005, 2009), is a well-known defender of a formal view of autonomy, and he motivates the view in large part because of its coherence with what he takes to be the best interpretation of liberalism. Normative capacity accounts, argues Christman, are in tension with liberalism. They seem, among other things, to suggest a "sliding-scale" (1988: 116) picture of anti-paternalist protections tailored to match the degree of decision-making competence and to invite state-sponsored perfectionist programs aimed at getting people's choices to align with the true and the good. These concerns are serious. Let me briefly explain why normative capacity accounts need not be illiberal.

As I have already suggested, the idea of responsible agency seems a crucial bulwark in any principled anti-paternalist case and normative capacity accounts seems better positioned than formal accounts to interpret what this sort of responsible agency comes to. Moreover, normative capacity accounts can, and I think should, accept the idea of negative autonomy rights. Most of us think (unjustified) paternalism *wrongs* people. We operate with the assumption that people have a *right* to decide for themselves in certain matters and that paternalism constitutes a usurpation of

their *rightful authority* to do so (Darwall 2006: 267-268). A negative autonomy right would protect a person's right (within suitable limits) to make bad self-regarding choices. This is just another way of saying that competent adults have powerful claims against others not to be interfered with in self-regarding matters. Since the right attaches to the capacity rather than its exercise, the right need not be thought of as conditional on making good choices.

One might, however, worry that even with negative autonomy rights in place, fidelity to the underlying normative structure would push normative capacity accounts toward three unpalatable conclusions: (i) significant scope-restrictions on who enjoys negative autonomy rights, (ii) variations in autonomy levels and therefore different autonomy rights for different competent individuals, and (iii) an invitation to make minute discriminations among persons concerning their normative competence. But this is not necessarily so.

First, we should distinguish between scalar and threshold assessments of autonomy. On a plausible view, negative autonomy rights attach to threshold normative competence. It is a further question where to set the threshold, but there is no reason to suppose normative capacity accounts are committed to setting it particularly high (cf. Griffin 2008: 156, Kauppinen 2011: 297). The lower the threshold is set, the less revisionary the account will be vis-à-vis standard liberal practice.

Second, variation above the threshold doesn't necessarily yield differential allocation of rights. To be sure, there is an important question about how to resist this conclusion. But the problem is more general and is familiar from discussions of equality. Egalitarians are committed to ignoring variation above some threshold, treating persons as equals even when they exhibit morally relevant properties to different degrees.[32] Hence, the problem is no worse for normative capacity theorists of personal autonomy than it is for egalitarians in general.

[32] Cf. Waldron's (2017: 84-127) discussion of equality in terms of the idea (originally from Rawls) of a "range property."

Third, it is open to normative capacity theorists to say that above some threshold of competence, treating persons equally requires what Ian Carter (2011) calls "opacity respect," that is, in such a way as to not make fine-grained distinctions about their normative competence. If so, then there would be moral reasons to desist from too closely tracking or using information about normative competence above the threshold, at least by certain agencies and within specified contexts (e.g. the state in relation to its citizens).

Fourth, a normative capacity account is consistent with thin procedurally defined operative standards in different domains. For example, one might think that a normative capacity account would demand a stingy approach to medical consent, e.g. in who gets deemed "capacitous." But this is not obvious. There may be good secondary practical and moral reasons for the existing standards, whether or not they adequately track normative competence. A variety of considerations—evaluative disagreement, proneness to error, liability to abuse, practical serviceability, and so on—speak in favor of thin procedural-looking proxy measures for normative competence which may, in practice, be over-inclusive from the point of view of genuine normative competence. Since the pressures of crafting realistic and well-justified policy may license and even require a departure from attempting to use genuine normative competence as criterial for the determination of negative autonomy rights in various setting, we must be cautious about attributing to normative capacity views pro-paternalist or illiberal policy implications in practice.

In short, normative capacity accounts need not be wildly revisionary vis-à-vis widely accepted liberal views about equality, rights, and respect. But do they commit us to perfectionist politics? The answer, I think, is that they do not. It is one question what autonomy is; it is a further question how autonomy is to be promoted—and by whom. Even if there are quite general agent-neutral reasons to promote anyone's autonomy, there may be good secondary reasons for insisting

that it is not everyone's business to promote everyone else's autonomy, and in particular, there may be special reasons to insist that states not be in the business of promoting autonomy. Whether the state may promote its citizens' autonomy is an important question, but it is orthogonal to the nature of autonomy. To see this, notice that it arises whether one adopts a normative capacity view or a formal view. Suppose, for example, that autonomy is, as Gerald Dworkin (1988: 20) maintains, "a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and to accept of attempt to change these in light of higher-order preferences and values." This type of view, too, might be combined with either perfectionist or anti-perfectionist political commitments. There is nothing about formal views of autonomy that commits those who hold such views to say to states or other agencies in authority, "Hands off on promoting this property!" A formal theorist might welcome state intervention in promoting autonomy, e.g., by promoting critical reflection. Conversely, there is nothing about normative capacity accounts that commits those who hold such views to say, "This property may (or should) be promoted by the state." Whether one has an inviting posture to state intervention is orthogonal to which view one takes about the nature of autonomy. The debate between liberal perfectionists and liberal anti-perfectionists—interesting and important though it is—should not drive our theorizing about personal autonomy.

Suppose, however, that normative capacity accounts do invite perfectionism in politics. Would this really be damning news? I think that is far from obvious. Some philosophers take it as virtually axiomatic for an account of autonomy of a liberal bent that it must respect neutrality and safeguard anti-perfectionism in politics (cf. Christman 2009, Dworkin 1988, Westlund 2009). But the question of how best to interpret the requirement of state-neutrality is notoriously complex and controversial. Proponents of formal theories all too often simply take for granted that liberalism

favors their view. Yet liberalism is a broad camp. There are sensible forms of perfectionist liberalism that have as good a claim as Rawlsian justificatory liberalism to being *bona fide* versions of liberalism (e.g., Green 1886, Hurka 1987, Mill 1859, Raz 1986, Sher 1997, Wall 1998).[33] To suggest that *all* substantive accounts of autonomy are illiberal won't work: normative capacity accounts, as I have suggested, need not have radically illiberal implications—and they do a good job interpreting the picture of responsible agency that seems to be presupposed by our liberal anti-paternalist practice. There is work to be done interpreting liberalism. At the very least, I think those who leverage anti-perfectionist arguments in favor of formal accounts of autonomy have more work to do in showing why we should antecedently favor non-perfectionist over perfectionist liberalism. And even if they make this case convincingly, it doesn't, so far as I can see, follow that autonomy is best interpreted in formal terms. For as I suggested in the last paragraph, what autonomy is and who gets to promote it are separate questions.

I conclude that normative capacity accounts need not necessarily conflict with liberal commitments. Much more, of course, would need to be said to allay fears that normative capacity accounts commit us to unattractive views of politics. What I have tried to show here, at least in brief outline, is that the conflict between liberalism and at least one variant of a substantive view of autonomy may not be as sharp or deep as sometimes supposed.

Conclusion

In this chapter, I have argued that formal views of autonomy face serious challenges. In particular, I have argued that they do not give us the right building blocks to make sense of the

[33] On Mill as a perfectionist liberal, see Brink (2013).

kind of responsibility that is plausibly at stake in autonomy, that they posit a fact/value asymmetry which creates an explanatory burden, and that they supply rather meager resources for helping us make sense of autonomy's normative role. I concluded by arguing that normative capacity accounts need not be on collision-course with liberal commitments. As I said at the outset, the choice between formal and normative capacity accounts of personal autonomy represents an important fork in the road which any theorist of personal autonomy must face. Many philosophers have bounded down the formal path, thinking it would take them in the right direction, but if the arguments in this chapter are along the right lines, it may be time to revisit the fork in the road and consider going the other way. The next chapter turns to the task of exploring the alternative.

Chapter 3

A Reason-First View of Autonomy

After introducing the challenge of conceptual pluralism in chapter 1, I argued for the plausibility of a research program associated with structured pluralism. While autonomy concepts are irreducibly plural, I argued, we should aim to fit them together into a theoretically coherent package. Chapter 2 mounted a case against formal views of autonomy. I argued that formal views face significant obstacles when attempting to make good on the full range of normative judgments associated with our thought about autonomy. In particular, I argued that formal views have difficulty capturing the idea that autonomy entails a robust sort of responsibility and plays a dual reason-giving role, being the sort of thing that calls for respect and is worthy of promotion.

This chapter develops a substantive account of personal autonomy. The guiding thought is that we can construct an attractive account of personal autonomy in terms of the idea that persons are normative agents characterized by capacities for discerning, and living in light of, genuine reasons and values. To accentuate the contrast with prevailing views of autonomy which privilege the idea of authenticity, I call this a reason-first view of personal autonomy. A reason-first view has the potential to address the challenges explored in earlier chapters. It yields a way of thinking about autonomy that makes good on judgments about responsibility and value, and it offers a promising strategy for achieving theoretical integration.

The plan for the chapter is as follows. Sections 1 and 2 begin by articulating the view's central presuppositions. Section 1 focuses on the reasons-responsiveness conception of responsibility, while section 2 clarifies the idea that there are substantive values and reasons which

can be said to constitute the aim of practical reason. Sections 3 and 4 articulate the view's core components. Section 3 explores the idea of rational control at the heart of the account, while section 4 shows how creativity and spontaneity can go together with rational control. Sections 5 through 7 then consider what resources the reason-first approach gives us for thinking about authenticity, independence of mind, and external independence, respectively.

1. Responsible Agency

At the heart of contemporary discussions of personal autonomy is a widely held picture of personhood and normative agency.[34] According to this picture, persons are capable of forming an evaluative conception of what matters in life. When they forge such a conception and succeed in living lives guided it, they are, in a modest but familiar sense, authors and creators of their lives.[35] Agents such as this, capable of evaluative and responsible self-fashioning, are thought to enjoy special dignity.[36]

An important point of division emerges, however, when we query the idea of responsibility at the heart of this consensus view. The above-described agents must be responsible agents. Because they can stand in the special sort of relation to their own lives in which they are authors and creators of those lives, persons must be, at least in some minimal sense, responsible for their lives.[37] But how is this responsibility to be understood? There are (at least) three interesting possibilities.

---

[34] Benson (1990, 2005); Berofsky (1995); Christman (1991); G. Dworkin (1988); R. Dworkin (1993); Feinberg (1986); Friedman (2003); Griffin (2008); Hurka (1987); Korsgaard (1996, 2009); Mele (1995); Meyers (1989); Oshana (2006); Rawls (1971/1999); Raz (1986); Watson (1975).
[35] Benn (1976: 125, 127), Dworkin, R. (1993: 224), Enoch (2017: 27), Griffin (2008: 150), Raz (1986: 369-370, 390).
[36] G. Dworkin (1988: 13), Griffin (2008: 152-153).
[37] Cf. Arneson (1980: 475), Buss (2012: 648), Dworkin, G. (1988: 20), Dworkin, R. (1999: 224), Friedman (2003, 21-22), Gewirth (1996, 115), Lucas (1966: 101, cited in Dworkin 1988: 6), Westlund (2009: 30-36).

The first is to interpret responsibility in terms of a deep-self conception, according to which persons are responsible for an episode of behavior if it reflects their fundamental commitments and orientation, quality of will, character, etc. Members within this family of views can be further distinguished by what Chandra Sripada calls their characteristic "demarcation moves," that is, their recipes for distinguishing features of an agent's psychology that count as part of the agent's deep self from those that do not count as part of the agent's deep self.[38] Identificationist variants require agents to self-reflexively identify with their own behavior through higher-order attitudes. This might be given a cognitivist interpretation, where identification proceeds by way of an agent's judgments (e.g., Watson 1975), or a volitional interpretation, where identification proceeds by way of an agent's conative states, like higher-order desires (e.g., Frankfurt 1971) or intentions (e.g., Bratman 1999, 2009). Non-identificationist variants, by contrast, do not require higher-order identification. The principle view of this kind is a sort of care-centered expressivism, according to which behavior reflects an agent's deep self if it expresses an aspect of her emotional profile—if it reveals what she actually cares about (e.g., Jaworska 2007, Sripada 2016).

The second possibility is to interpret responsibility in terms of a reasons-responsiveness conception. According to philosophers working in this camp, responsibility requires the ability to appreciate and conform to the norms that apply to one's conduct (Brink & Nelkin 2013, Fischer and Ravizza 1998, Nelkin 2011, Vargas 2013a, Wallace 1994, Wolf 1990). Very roughly, one will be a responsible *agent* if one is the kind of agent who can recognize and respond to normative reasons, and one will be responsible for some bit of behavior on a particular occasion if one is able to recognize and respond to the reasons in play on that occasion. Although details differ from one author to the next, the core idea behind such views is that responsibility requires a kind of rational

[38] Sripada, "At the Center of Agency, The Deep Self" (*manuscript*). For the following taxonomy, I draw on Sripada (ibid) and on Wallace (2014).

control. Since behavior is assessed relative to normative standards, it is the ability to control one's behavior in light of these standards that matters. If one is to be responsible for whether one's behavior conforms to normative standards, one must possess abilities to apprehend and comply with the relevant standards.

The third possibility is to interpret responsibility in hybrid terms. Following Gary Watson (1996), it is useful to distinguish two senses of responsibility. In the *attributability* sense, persons are responsible for acts or attitudes insofar as those acts or attitudes reflect something about their quality of will or character. In the *accountability* sense, persons are responsible for acts or attitudes insofar as they can appropriately be held liable to a range of sanctioning responses, like blame and punishment. One might deny that there are these different senses of responsibility, or one might accept that there are these different senses but insist on understanding what they come to in terms of a uniform theoretical construct. However, a more plausible alternative is that the two senses are genuinely distinct and that deep-self and reasons-responsiveness views of responsibility are specifications of the conditions making these different forms of assessment appropriate. Deep-self views give us a story of when behavior reflects significant morally assessible aspects of persons and can be appropriately attributed to them, reflecting well or poorly on what they are like; reasons-responsiveness views give us a story of when agents are appropriately held to account for some bit of behavior.

Formal views of autonomy are committed, either explicitly or implicitly, to interpreting the kind of responsibility at issue in personal autonomy in terms of a deep-self conception. As we saw in the last chapter, formal views of autonomy are committed to *not* making capacities of reasons-responsiveness criterial of autonomy. It follows that they cannot understand the responsibility that is involved in autonomous agency in terms of responsiveness to reasons. Instead, formal views of

autonomy must understand responsibility in terms of *authenticity*, i.e., conditions which specify when some act or attitude is an agent's own. Consider two quotes, one by Gerald Dworkin, the other by Marilyn Friedman:

> [A]utonomy is … a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values. By exercising such a capacity, persons define their nature, give meaning and coherence to their lives, and take responsibility for the kind of person they are. (G. Dworkin 1988: 20)

> [I]n the last analysis, what matters to someone, what she self-reflectively cares about, when effective in and reflected in her action, makes her behavior autonomous. (Friedman 2003: 8-9)

The basic idea, elegantly encapsulated by Dworkin and Friedman, is that autonomy consists in the exercise by persons of their capacities for identifying with, and living from, an evaluative conception. Crucially, however, there is no requirement in these formulations that autonomous agents track any genuine values or principles. Nor is there any aim internal to autonomy that enjoins agents to try to get matters right. Fundamentally, the question agents ask is not, "Is this value or principle correct?" but, "Does it meet with my reflective approval?" To be sure, there is an important critical moment in the life of autonomous agents: they must be prepared to critically reflect on their attitudes, values, and principles; they must be ready to subject their evaluative conception to scrutiny and jettison it if needed. Moreover, such reflective self-vetting must occur under conditions of what Dworkin (1988) calls "procedural independence." But these reflection conditions and their procedural safeguards are not, on Dworkin and Friedman's views, designed to ensure responsiveness to genuine normative truths. Rather, they are designed to ensure that the evaluative conception a person is guided by is *her own* in some suitably deep sense.

On classic views, like Dworkin's and Friedman's, authenticity comes about through an essentially identificationist route. Persons *find* themselves with desires, proclivities, cares, values, and so on. Reflective self-audit then turns the mere givenness of these elements of one's psychology into desires, proclivities, care, values, etc., *of one's own*. Other autonomy theorists have described authenticity mechanisms that do not require reflective identification. They maintain that agents are identified with aspects of their psychology by some other means, for example, because their attitudes (i) are such that agents would not disavow those attitudes if they became aware of their source (Christman 1991a), (ii) are such that agents would in principle be prepared to answer for those them (Westlund 2003, 2009), (iii) are in accord with the ground-level normative policies which define agents over time (Bratman 2009), or cohere with an agent's other attitudes (Waddell Ekstrom 2005). In any case, however exactly persons become identified with aspects of their psychology, it is a deep-self mechanism that constitutes persons as responsible for their lives and choices. If persons are special creatures who can forge and live from an evaluative conception and thereby be responsible for their lives, on formal views of autonomy this responsibility ultimately has to be understood in terms of those lives reflecting or expressing a person's deep self.

As I argued in the last chapter, it is questionable whether this notion of responsibility suffices for an account of autonomy. If that's right, it leaves two options. Either a notion of responsibility adequate to our thinking about autonomy will be a purely reasons-responsiveness conception or it will be hybrid. Which of these options is ultimately more plausible depends on complex background assumptions, which I lack space to pursue. For my part, the hybrid option seems more plausible. The reasoning is straightforward. First, I think a hybrid conception is already necessary to capture the full range of *moral* responsibility judgments, so there is reduced

theoretical motivation to keep the kind of responsibility involved in autonomy "pure." Second, I argued in chapter one that authenticity is *part* of our thinking about autonomy, in chapter two that formal theories appear to give us resources for thinking about this dimension of autonomy, and in this chapter that formal theories of autonomy reflect a deep-self conception of responsibility. It is therefore a plausible hypothesis that a deep-self conception of responsibility plays some role in our thinking about autonomy as well. Third, it seems to me likely that some ways of using the language of autonomy may come apart from reasons-responsiveness. Just as in the moral case it seems plausible that persons may be morally assessible for their quality of will or character without being reasons-responsive, in the case of personal autonomy it seems plausible that there are forms of respect owed to persons *qua authentic agents*, i.e. agents who have crystalized a particular identity or perspective on the world, that is genuinely their own, yet who are not (or not very) reason-responsive. Such agents will not be paradigms of autonomy on the kind of view I develop here. Nor will they, if they do not enjoy threshold-level reasons-responsiveness, qualify for strong anti-paternalist protections. Nevertheless, we can recognize that such agents have forged, and are successfully living from, an independent perspective on the world. This alone suffices for a kind of recognition respect. It will not be the kind of recognition respect that tracks their status as competent choosers but one which acknowledges their status as subjects with a view of their own.

For my purposes, the following two theses are more pressing:

Indispensability. A complete and satisfactory account of autonomy cannot do without reasons-responsiveness.

Priority. Reasons-responsive responsibility is more fundamental than deep-self responsibility in an adequate account of personal autonomy.

The first claim is needed to vindicate the full range of normative commitments associated with our thinking about autonomy, as shown by the argument of chapter two. It is key to this chapter because I build my account of autonomy around the core idea of reasons-responsiveness. The second claim is not strictly necessary but reflects the general drift of the account developed here. If that account is correct, reasons-responsiveness turns out to have priority in two ways. First, reasons-responsiveness is *explanatorily fundamental* because it is used to integrate and make sense of other ideas. Second, and related, there is at least a moderate *material dependence* (i.e., one that is substantive and non-conceptual) between reasons-responsiveness and authenticity insofar as the former secures the latter but not vice versa.

I said above that the core idea behind reasons-responsive views of moral responsibility is that responsibility requires a kind of rational control. In many ways, this is my guiding thought as well. While the kinds of reasons and values at play in personal autonomy is not limited to moral considerations, the idea of rational control vis-à-vis operative norms is the same: there are substantive norms, grounded in reasons and values, that apply to one's behavior, and one's behavior will display more rational control when it is appropriately sensitive to these norms. Having said this, the notion of rational control by itself doesn't capture everything. That is because there can be choice that is unconstrained or underdetermined by reasons. Autonomy makes room for spontaneity, creativity, and discretionary choice. I therefore distinguish *rational control* from *elective control*, where the former is understood in terms of sensitivity to reasons and the latter is the discretionary freedom that is left over when reasons run out. As I explain further below, elective control will be a genuine kind of *control* insofar as it operates within the bounds of rational control.

2. Substantive Value and the Aim of Practical Reason

The reason-first model makes significant assumptions. In particular, it assumes (i) that there are substantive values and reasons to be tracked, and (ii) that, at least in some significant sense, the point of practical reason is to "get it right," that is, to make agents responsive to these values and reasons. While I cannot entirely vindicate these assumptions, I want to at least briefly indicate why I think they are credible and, in many ways, quite modest.

(i) The idea that there are substantive truths about what agents ought to do (or what it would be good for them to do) is one which is widely shared in contemporary philosophy, including by those who give formal accounts of autonomy. It seems difficult to deny that there are genuine practical reasons.[39] These reasons seem to come in two kinds, moral and prudential. By all appearances, each of these domains contains substantive normative truths, and what agents ought to do is some function of these substantive truths, indexed to their situation. The idea of practical reason, and of rationality more generally, requires such an independent normative standard. The idea of rationality is normative, not descriptive: it supplies a standard whereby actual behavior can be critically assessed. It follows that actual behavior doesn't determine the standard, hence, that the standard must be independent of actual behavior. Short of giving up on the idea of rationality in general, and rationality applied to action in particular, it is difficult to see how one could give up on the idea that there are normative standards governing behavior.

There is, to be sure, disagreement at two levels. The first concerns *intra*-normative structure. For example, some philosophers take the distinction above, between moral and prudential reasons, to be quite deep; others deny this. Some philosophers take The Right to be determined by The Good; others deny this. And so on. The second concerns *meta*-normative

[39] The operative notion here is that of a *normative* reason, not that of a *motivating* reason. The former are considerations that set the standard for conduct; the latter are considerations that explain what agents actually do.

commitments. Some philosophers think values reduce to reasons; others deny this. Some philosophers take reasons to be a function of an agent's motivational set; others deny this. And so on. But regardless of how one comes down on these questions, if one takes seriously the above idea of rationality *as a normative concept*, then there remains the possibility of a gap between occurrent behavior and normative standard. In other words, as long as one is prepared to hang on to the idea of rationality as a normative concept, one can countenance the idea of normative reasons. What precise content those reasons have, how they relate to each other, and how they ground in facts about oneself, will vary depending on one's intra-normative and meta-normative views. But the reason-first approach doesn't dictate any particular interpretation of these matters; it only requires *that there be* such reasons applying to the conduct of agents. This is a fairly modest claim and it is compatible with a wide range of views about both the ground and content of the reasons that apply to agents.

Take prudential reasons. The reason-first approach maintains that, to count as autonomous, the project of forging and living from an evaluative conception cannot float free from the things that actually matter. Some of what matters concerns an agent's own good. According to the reason-first view, then, forging an evaluative conception and making autonomous choices with respect to one's own life, requires adequate sensitivity to the facts that constitute one's own good. These facts might be interpreted in terms of any of the leading views of welfare: hedonist, desire-satisfaction, objective list, or perfectionist. To get off the ground, all the reason-first approach requires is the thought that there are normative standards for choice which are independent of an agent's occurrent desires, preferences, goals, etc. In other words, there must be normative standards independent of an agent's momentary whim. Such normative standards are countenanced by all of the above views of welfare—even desire-satisfaction views. Any plausible

version of a desire-satisfaction view introduces mechanisms of idealization, so that the preferences that matter for purposes of determining an agent's welfare are indexed, not to what an agent *actually happens* to want, but to what she *would want* were she adequately informed (Sobel 2009). Objectivity of normative standards, in other words, is consistent with views assigning a larger or smaller role to the subjective states of agents in determining what it is rational for them to do.

Of course, the reason-first approach to autonomy makes no sense if there aren't normative facts for the domain of choice which make it the case that choices can be better and worse, right and wrong, wise and unwise. But to think there are such facts is not a particularly radical supposition; it is the *denial* of that supposition that seems radical. Short of dispensing with the idea of (practical) rationality altogether, one will have to be prepared to countenance gaps between behavior and the normative standards that apply to that behavior.

(ii) The idea that practical reason has a "point" or "aim" may bring to mind a variety of special theses about action and practical agency which are quite controversial, theses like the "guise of the good thesis" or the idea that there is a constitutive aim of action. I wish to avoid these special theses. I offer two interpretations of "aim-talk," one deflationary, the other non-deflationary, which I think are plausible and should not be all that controversial.

The deflationary interpretation is this. Practical reason, I take it, is the human capacity to rationally settle practical questions. Often, this proceeds by way of practical reason*ing*, i.e., through conscious deliberation about the merits of competing options. But not always. The human capacity for practical reason is manifested in a sensitivity (only sometimes reflective) to a variety of reasons relevant to behavior. If practical reason *just is* this sensitivity to reasons, then saying that the "aim" of practical reason is tracking reasons is just a re-description of capacity-talk. In this

sense, clocks "aim" to tell time and radio telescopes "aim" to capture electromagnetic signals coming from outer space with certain wavelengths and frequencies.

The non-deflationary interpretation is that *persons* have, and are committed to having, the aim of practical reason, i.e., they have, and are committed to having, the aim of exercising their capacities for practical reason. Part of this claim rests on empirical generalization. Trying to figure out what it is one ought to do, weighing reasons in deliberation to this end, entertaining justificatory demands from actual or notional others—these are familiar and ubiquitous features of the human experience. It is plausible that this is so because most people are *in fact* motivated to try to figure out what they have reason to do. Among agents capable of practical reason, it is fair to say that most care about doing what it is they have reason to do.

The other part of the claim rests on an interpretation of what persons are rationally committed to. The best way to see this is in terms of their being valuers, or at any rate, valuers of a certain sort. Samuel Scheffler offers an illuminating account of valuing. According to Scheffler (2011: 32), human valuing is a complex cognitive and affective syndrome involving the following range of attitudes and dispositions:

1. A belief that X is good or valuable or worthy
2. A susceptibility to experience a range of context-dependent emotions regarding X
3. A disposition to experience these emotions as being merited or appropriate
4. A disposition to treat certain kinds of X-related considerations as reasons for action in relevant deliberative contexts

The language of valuing is, of course, sometimes used more loosely. Dogs may be said to value their food, young children their play, and persons who have lost (or never developed) certain cognitive capacities may be said to value all sorts of things in the sense that they *care* about them,

that is, in the sense that they display characteristic patters of emotion (Jaworska 2007, Sripada 2016). But Scheffler's analysis helpfully captures a more restricted notion of what we might call *characteristically human valuing*. The valuing that normal adult humans engage in is a complex phenomenon. It includes object-directed care, but it also recruits an evaluative or normative perspective on the object of care. In addition to being susceptible to a range of emotions, characteristically human valuing includes believing that the object is *valuable*, that it *merits* certain emotions, *calls* for certain actions, and so on.

On this picture, characteristically human valuing is suffused with normative commitments. These normative commitments implicit in valuing bring rational pressure with them, e.g., to entertain justificatory demands by oneself or others. One can have peculiar or parochial values, but it strains coherence to imagine having values one finds inappropriate and unjustifiable or for which one wouldn't be prepared (at least in principle) to "go to bat." Notice the tight conceptual connection between values and reasons. On the one hand, it is hard to get a grip on *practical* reasons (that is, *normative* reasons which are practical) which don't realize some kind of value (Raz 1999). On the other, to say that some X has value is to say that X gives agents reasons of various kinds (Scanlon 1998). The concept of a (normative) reason just is the concept of a rational "favorer" which contributes to justifying a response of some kind. Hence, insofar as characteristically human valuing embeds claims about value, it comes with claims about reasons, and claims about reasons are tied to justification.

We can approach the idea of rational constraints from another angle. Valuing includes not only believing *valuable* but *believing* valuable. The truth-directed character of belief sets the central dynamic of believing and places rational constrains on what it is possible to believe. It is difficult to adopt beliefs one thinks are false; if this is possible at all, it is a quite marginal

phenomenon, and one which has to contend with the fact that belief is a state which purports to represent how things are, which in some sense "aims" at truth (Velleman 2000). To the extent that one believes the principles and values one endorses, one's commitment inherits rational pressure from the truth-aim of belief.

Insofar as persons engage in distinctively human valuing, then, they inherit rational commitments in virtue of being valuers and believers:

1. Distinctively human valuing manifests not only in a pattern of object-directed concern but also in a kind of meta-concern for the appropriateness of the valuing response to the valued object.
2. Being in the business of believer saddles one with rational commitments relating to the adequacy of one's beliefs.

These claims have significant implications for how persons' selection and retention of principles and values is best modeled. The authenticity-centered perspective that dominates thinking about autonomy does not adequately capture what people actually tend to care about or what they are rationally committed to caring about. Take the reflective identificationism of Dworkin and Friedman referred to earlier. On their view, one reflects on and then chooses to stand behind one's principles and values, or else jettisons them for new ones. There are, of course, rational constraints on this process—otherwise it would be difficult to recognize it as one that is supposed to characterize the activity of rational agents. One must, at a minimum, recognize norms of thin procedural rationality, like consistency constraints on beliefs and intentions. But these are not very substantively demanding norms: they leave a great deal of room for free play. And that makes the choice of principles and values look altogether too arbitrary. It makes it seem like, so long as persons engage in the right reflective procedure, they can rather whimsically select the principles

and values that are to their liking—the one's they happen to identify with. But agents' identification is a fact about their psychology, not a fact about the adequacy of their values and commitments. Identification can be substantively misguided. Persons recognize this fact and typically care about it. They do not take their own endorsement as a badge of success. Nor is this general empirical fact about persons entirely accidental: it is grounded in commitments internal to their normative agency. Qua valuers and qua believers, persons are committed to standards of adequacy. Internal to the viewpoint of practical agency, then, persons are committed to operating with correct or adequate values and norms—not just *any* norms and values one happens to want or to find oneself with.

This rational dynamic means that practically rational agents are not well-described as happily adopting values and principles irrespective of their sense of the normative adequacy of those values and principles. They might, of course, get it wrong, but they do not *set out to get it wrong*—or at any rate, if they do, they have to fight the central commitment dynamic which is internally governed by standards of rational adequacy. The kind of whimsical value-adoption envisioned by (pure) authenticity models of autonomy does not capture very well what people are actually like or must be like as rational valuing agents.

The reason-first approach delivers a view of autonomy which does not sever it in this way from the aim of practical reason. It speaks precisely to the central thing persons are up to as practical agents, to their project of living in ways they have reason to. Far from being a costly assumption, then, the idea that practical reason has an aim can be understood as an independently plausible claim that puts pressure on formal views and beckons for an alternative.

3. Rational Control

At the heart of the reason-first view of autonomy is the idea of rational control. The basic idea can be put simply:

Core Formula: All else equal, a person is more autonomous, the more rational control (reasons-responsiveness) she enjoys.

The core formula says that autonomy is directly proportional to rational control. Rational control, in turn, is interpreted in terms of responsiveness to reasons.

Accounts of rational control in terms of responsiveness to reasons have their origin in theories of moral responsibility, but the basic idea can be generalized. It is this: for any normative domain in which agents are responsible vis-à-vis the content of that domain (in a sense sufficiently robust to ground accountability responses), they must have abilities to register and conform to that content. There are important questions about how the concept of reasons relates to other normative concepts and about whether any particular normative concept is most basic. For our purposes, what matters is only that (i) the concept of reasons is among the fundamental normative concepts, (ii) normative claims can be expressed in terms of reasons, and (iii) genuine normativity (as opposed to merely norm-governed institutions and conventions) is characterized by the presence of genuine reasons. Assuming one can express any genuine normative claim in the coin of reasons, reasons are a handy all-purpose way of speaking about an agent's relationship to genuine normative domains.

To conceptualize rational control in terms of reasons-responsiveness, we can make two basic distinctions. Following Brink and Nelkin (2013), we can distinguish, first, between

normative competence and *situational control*, where the former consists in the more or less cross-situationally stable ability to recognize and conform to relevant norms and the latter consists in situation-indexed ability to do so on a given occasion. Second, the ability to recognize and conform to relevant norms can be thought of as involving *cognitive* and *volitional* ingredients. Brink and Nelkin (2013) think of these ingredients as, in the first instance, aspects of normative competence, but it is part of their view that normative competence interacts with features of an agent's circumstances to yield more fine-grained, situation-indexed abilities. Hence, on their view, it is just as appropriate to distinguish cognitive and volitional aspects of situational control.

The challenge in the rest of this section is to apply the idea of rational control as reasons-responsiveness in the context of thinking about personal autonomy. This raises a host of questions. For example, what is the relevant domain of reasons? What is the relationship between these reasons and one's evidence? Must one actually exercise one's capacities for reasons-responsiveness or merely possess them? These questions interact with others. For example, what are the relevant range of attitudes with respect to which individuals must enjoy rational control? Let's consider these and other questions in an effort to help refine the view.

Binary vs. scalar.

According to the core formula, autonomy and reasons-responsiveness are both scalar phenomena, varying in a continuous manner. However, we need to distinguish two different notions of autonomy, one scalar, the other binary (cf. Brink and Nelkin 2013). The binary notion is important because (as with freedom and responsibility) we have to be able to make all-things-considered judgments that track a *sufficiency threshold*. To be autonomous in this binary sense

("autonomous full-stop"), one needs to enjoy *sufficient* rational control. Moreover, it is plausible that below some threshold of rational control, persons are not autonomous at all—not even a little. If autonomy only "kicks in" above a threshold-level of reasons-responsiveness, then we can think of the binary notion of autonomy as useful for purposes of tracking that threshold, whereas the scalar notion is useful for tracking variation above the threshold.

Domain

What exactly is the domain of reasons relevant to personal autonomy? In the case of moral responsibility, the answer seems relatively clear: moral reasons. In the case of personal autonomy, the answer is less clear. As a first approximation, we might say that the relevant reasons are practical reasons, where this includes both *moral* and *prudential* reasons. Think of decisions about how one is to live one's life. These decisions will be underwritten by appeal to values and principles that have moral and prudential content—what it is to live a good life, what sorts of things are worth pursuing, how we are to treat others, etc.

One might think that some of these considerations about worth and excellence have a quasi-aesthetic character. Whether or not they do, there are other considerations that clearly do. It seems possible not only to autonomously decide to attend a particular symphony or pursue a life in art, but also to autonomously make aesthetic judgments, e.g., about the merits of the symphony. The status of aesthetic norms is controversial. Perhaps they are mere matters of taste. If so, they will be person-relative, like many ordinary preferences. If aesthetic norms are more objective, however, then to the extent that individuals make judgments and choices involving such norms, they would need to have facility with relevant *aesthetic* reasons (cf. Nelkin, *forthcoming*).

More importantly for our purposes, however, beliefs in general matter to personal autonomy. They matter both as states with respect to which persons can count as autonomous, i.e., people can autonomously believe. They also matter as states which help constitute persons as autonomous overall, i.e. people can be overall autonomous (in part) in virtue of their beliefs. These twin roles of belief will be explored further below. For now, it is enough to note that on the picture of autonomy developed here, rational control will also require appropriate facility with *epistemic* reasons.

Focus

Judgments about autonomy may differ in focus. They may focus on persons, on actions, or on attitudes. Ultimately, these different targets of assessment are closely related. The core formula is put in terms of a person-level property. It says that persons are more autonomous the more rational control they enjoy. But such assessment depends on persons enjoying rational control in respect of actions and attitudes. This parallels the way we make assessments of other kinds. We say that persons are rational *and* that particular actions and attitudes of theirs are rational. Moreover, the rationality of persons is some function of the rationality of their actions and attitudes. Persons cannot be rational if none of their actions or attitudes are rational.

Some accounts of autonomy employ a choice-focused paradigm. Recall the quote from Dworkin: "[A]utonomy is … a second-order capacity of persons to reflect critically upon their first-order preferences, desires, wishes, and so forth and the capacity to accept or attempt to change these in light of higher-order preferences and values." Accounts like these import the focus on choice to thinking about what it is to be autonomous in respect of an attitude. The agent endorses

an attitude, thereby choosing it and making it her own. But this is a dubious model for the autonomy of at least some attitudes. This is clearest in the case of belief. Belief represents its content, propositional content p, *as* true. It is therefore typically sensitive to evidence bearing on the truth of p. And this seems entirely apposite: belief *should* be sensitive to such evidence. It is clear neither what it would mean to endorse a belief, nor clear that, whatever it means, endorsement is the right sort of thing to render belief autonomous. So choice-centered endorsement accounts offer a bad model of what it is to be autonomous in respect of at least some attitudes. I'll return in a moment to the question which attitudes matter for autonomy—and whether beliefs are among them.

The reason-first approach gives choice its due. As I'll argue below, it makes room for choice at a variety of junctures. But since it has an alternative and fundamentally different recipe for guiding assessments of autonomy, it need not generalize the choice paradigm. What we should be interested in, says the core formula, is whether a person has rational control. If and when there is a choice to be made, it is a person's rational control vis-à-vis that choice that determines the extent of her autonomy with respect to it.

Types of Attitude

Much of the literature since Frankfurt focuses on the importance of desires and preferences. This focus is too narrow. On a psychologically realistic picture, our practical lives are mediated by many different structures: by plans, intentions, policies, ambitions; by concerns, cares, loves, loyalties, commitments; by beliefs, hopes, fears; and so on. In some cases, these structures are simple attitudes, in others, complex ensembles of attitudes. Within the rational control framework,

any type of attitude may in principle be relevant to assessing a person's autonomy. Such attitudes may include *practical attitudes* like desires and intentions, *doxastic attitudes* like beliefs, and *affective states* like emotions. Attitudes of each kind can drive, influence, and modulate attitudes of the other kind and issue in behavior. What matters on the rational control picture is that persons are responsive to reasons. This is a high-level property, which is subserved by psychological mechanisms of various kinds. When it comes to conceptually parceling up the mind, there is no need to focus narrowly on just one kind of state or structure to the exclusion of others.

Views of autonomy premised on the idea of higher-order reflective endorsement, by contrast, are committed to privileging certain aspects of the mind over others. Take Dworkin's version, which appears to hold fixed "higher-order preferences and values" in light of which people choose. Such privileging raises awkward questions (cf. Watson 1975). Are people simply stuck with their higher-order preferences and values? What about their beliefs, with which their values and higher-order preferences are inevitably bound up? Can *beliefs* be autonomous? And if people are non-autonomous in respect of their beliefs, values, and higher-order preferences, how could they become autonomous merely in virtue of the fact that they endorse their own attitudes via endorsement-structures composed of such non-autonomous attitudes?

Since endorsement accounts like Dworkin's make the choice paradigm central, they have, as I remarked above, difficulty making sense of what it is for beliefs to be autonomous. Might there be principled reasons for excluding beliefs as attitudes with respect to which persons can be autonomous? I don't see any. On the contrary, it seems to me there are principled reasons for including beliefs within the repertoire of attitudes that are relevant to assessments of autonomy. First, it seems both normative and non-normative beliefs matter to autonomy. Normative beliefs matter because, as we have seen, they are tied to what people value and, hence, are partly

constitutive of an individual's evaluative perspective. Non-normative beliefs matter because, as I noted in the last chapter, it is hard to take seriously a view of autonomy on which it can float entirely free from adequate information about the world. Second, when it comes to determining how an agent behaves, mind-to-world states like belief and world-to-mind states like desire are equally momentous, and each type of state only gets its functional role in conjunction with sates of the other type. It seems arbitrary to focus on only one type of state to the exclusion of the other.

The idea that one might be autonomous in respect of beliefs fits nicely with the reason-first view. Philosophers who have taken the idea of autonomy or freedom in respect of beliefs seriously have typically privileged the idea of reasons-responsiveness (McHugh 2017, Pettit and Smith 1996, Raz 1997). One might leverage this as further reason to go for a rational control view. Assuming we need an account of autonomy on which people's being autonomous in respect of beliefs is an intelligible possibility, and assuming explanatory simplicity is preferable all else equal, this favors views which can accommodate the possibility of autonomous belief with the fewest ad hoc assumptions (cf. Sayre-McCord & Smith 2014).

If the core formula is right, the recipe of reasons-responsiveness is quite general, covering not only belief but other attitudes and psychological states as well. Since what matters, according to the control model, is the high-level property of reasons-responsiveness, that model can be quite flexible about which attitudes matter to assessments of autonomy. We need not focus narrowly on any particular mechanism, structure, or activity.

Kinds of control

Rational control has to do with one's fitness for normatively assessed behavior, either in general or on some occasion. When behavior is judged according to a normative standard, the key question is: does the individual have the cognitive and volitional wherewithal to recognize and comply with the normative standard? It is worth clarifying how this type of control relates to others.

(a) Performance control has to do with one's power to effectively bring about behavioral goals: to raise one's arm, to shoot a hoop, to remain calm while the spider crawls over one's hand. Such control is best thought of in terms of reliability in bringing about a goal state across some range of circumstances (Shepherd 2014). Performance control does not guarantee rational control. One might enjoy high performance control in one's engagement in an activity which one has decisive reasons not to engage in but have little capacity to register or act in light of that normative fact. Conversely, however, rational control does entail adequate performance control: to be able to respond to reasons requires being able to adequately regulate conduct in light of those reasons, i.e., to exercise appropriate performance control.

(b) Proximal control has to do with whether one can bring about some goal at will. Consider the following familiar contrast. You can raise your arm at will, but you cannot choose to believe something for which you lack evidence or change an emotion at will. However, even though you cannot alter your beliefs and emotions at will, you are not entirely powerless to alter them. To alter beliefs, you might change your environment, choose to deliberate, gather more evidence, cultivate a frame of mind, and so on. To alter emotions, you might leave the room to "cool off," change the music, hug a friend, reframe an experience, remind yourself what you really care about, and so on. The intuitive contrast here is between direct and indirect control, i.e., with whether some outcome or upshot can be brought about by an act of will.

The distinction between direct and indirect control is orthogonal to performance control and rational control. One might enjoy indirect though high performance control, as when one reliably brings it about that one believes the light is on by flipping the light switch. Or one might enjoy direct but low performance control, as when one fails to reliably bring it about that one stands up on ice. Likewise, one might enjoy indirect though high rational control, as when one reliably discerns reasons through extended deliberation and then conforms to these reasons. Or one might enjoy direct but low rational control, as when one perceives and intends to conform to reasons but is impaired by depression from conforming to them.

(c) Option control has to do with one's opportunities. Reza is passionate about cello, but he has neither money for lessons nor time for practice. Ashraf is not particularly passionate about cello, but she can afford lessons from the best and has plenty of time to practice. Reza lacks a kind of control that Ashraf has. Even if both of their musical talents languish, the explanation will differ. Ashraf would prefer to do something else with her time. Reza's "hands are tied."

Option control is not required for performance control. In Raz's (1986) example, the woman who is trapped on an island with a hungry beast and spends every waking moment trying to outsmart the predator has few meaningful options. Nevertheless, she may have high performance control in this activity. Being much smarter than the beast, she might be able to consistently elude it.

The relationship between option control and rational control is more complicated. Whether, and in what way, option control matters to rational control depends on the contexts and purposes of evaluation. I argued in the last chapter that having options plausibly matters to autonomy. But now that we are working out what it is to be autonomous in terms of the idea of rational control, we can be a bit more nuanced. When it comes to living an autonomous life, options are plausibly

required. Hence, the woman trapped on the island will not be living an autonomous life. She is forced by dint of circumstance to live the life that she lives. When it comes to having attitudes with respect to which one counts as autonomous, however, options are plausibly not required. The woman has highly rational beliefs, desires, and emotions. She perceives what there is reason to believe and do and behaves accordingly. Rational control, then, does not always require options.

The notion of rational control is broader than the notion of option control. In some contexts, option control matters for rational control; in others, it does not. Which contexts are relevant? I have no precise rule, but roughly, options matter within the *sphere of choice*. It is crucial to the reason-first perspective that not everything must be chosen (or endorsed). In particular, choice is not required to make attitudes autonomous. But there clearly *are* things we choose (or might be in a position to choose) in life, like friends and associations, vocational and avocational pursuits, and so on. When choice matters, so do options. Hence, if Reza lacks opportunities to play cello, this *is* a limit on his rational control—at least, in respect of that choice. But he may be cognitively undeluded about his situation and respond with all the appropriate emotions. In this sense, of course, he suffers no deficit in rational control.

Objectivity

Responsiveness to reasons might be understood in more or less objective senses:

- Responsiveness to reasons *as we see them*
- Responsiveness to reasons *as indicated by our evidence*
- Responsiveness to reasons *as they are*

Responsiveness to reasons *as we see them* is too weak for the kind of rational control we are after, though it may be one way to understand what enjoying independence of mind amounts to (more on this below). This leaves open whether responsiveness to reasons should be understood in terms of the second or third claim. Some philosophers proposing rational control accounts of autonomy (or freedom) say things that suggest they think the relevant notion of reasons-responsiveness is evidence-relative (Raz 1997: 223-224; Smith & Sayer-McCord 2014: 149, footnote 23; McHugh 2017: 2752). Others say things that suggest they think the relevant notion of reasons-responsiveness is fact-relative (cf. McDowell 2010).

The difficult question is about how we should understand the objectivity of *practical* reasons. It seems clear enough that reasons *for belief* are evidence-relative. I have reasons to believe p, even if my evidence for p is misleading. If my reasons for belief were fact-relative, I would presumably have reasons to believe all and only true claims, perhaps indexed to what is (in a way that would need to be filled out) accessible to me. But then I would have reasons to believe many things for which I now lack evidence. By contrast, it is not at all clear practical reasons are evidence-relative rather than fact-relative. There are plausible views on which practical reasons can be understood as evidence-relative or fact-relative (Graham 2010, Zimmerman 2014). Which of these is correct? I cannot hope to settle this question here. I'll therefore simply say what seems plausible to me. Luckily, the reason-first approach can afford to be ecumenical on this issue, so what I say isn't a prerequisite for signing on to the broader view.

In general, it seems to me a mistake to insist that practical reasons are either fact-relative or evidence-relative. We need both concepts, and we need them for different purposes. For example, we need the evidence-relative concept to make sense of a why a doctor who prescribes medicine against her evidence would be negligent even if it turned out to be the most effective

drug. On the other hand, we need the fact-relative concept to make sense of what it is the doctor ought to do if she had full evidence, and what other persons, better apprised of the facts, ought to counsel her to do ("I realize you *think* you ought to administer medicine A, but you *really* ought to administer B").

The same goes when thinking about autonomy. There is no simple answer about which of these concepts is the relevant one. Return to the examples from last chapter, due to Al Mele (1995: 179-181). Connie chooses an investment plan but is deceived by the company offering the plans. King George rules his kingdom but his decrees are based on information that has been cleverly curated by his ministers to achieve their own designs. Aren't these characters seriously defective in autonomy? On the other hand, however, Connie and King George have this much going for them: they are responding to the evidence they have. They are not engaging in wishful thinking, making decisions in ways that look erratic or blatantly irrational, and so on. Rather, it appears they are acting quite rationally on the basis of information available to them. Isn't there then some obvious sense in which they are acting quite autonomously after all?

The answer to both questions is "yes," and we need the different concepts of reasons-responsiveness to explain why. Connie and King George are reasons-responsive in respect of their beliefs. That is, their beliefs are responsive to the reasons they have for believing. Hence, they are fully autonomous in respect of these attitudes. But their beliefs turn out to be false and lead them to do things that (let's assume) they have fact-relative practical reasons not to do, like make bad investments or ruin a kingdom. If all practical reasons are interpreted as evidence-relative, then Connie and King George are through-and-through responsive to reasons, hence suffer no impairment or diminution of autonomy. Intuitively, as I argued in the last chapter, this is not credible. We might adopt the counterintuitive interpretation that Connie and King George are

perfectly autonomous if we had very strong theoretical motivation for thinking that all practical reasons must be evidence-relative. So far as I can see, however, we lack such motivation. In general, I suggested above, we need *both* concepts of reasons. Hence, the evidence-relative interpretation of reasons-responsiveness does not force itself upon us for independent theoretical reasons.

Moreover, there are considerations internal to our thinking about autonomy that pull the other way. First, there is the intuition that Connie and King George suffer some impairment of autonomy. Second, there are theoretical considerations to support and make sense of this intuition. In the last chapter, I invoked ideas about responsibility and normative adequacy to suggest that autonomy requires responsiveness to *genuine* reasons. Both ideas are best supported by a fact-relative interpretation of responsiveness to reasons: to track the reasons there genuinely are is to be more richly responsible and to enjoy a kind of agency that is more worthy of promotion. And in this chapter, I invoked the additional idea of an aim internal to practical reason. I argued above that people tend in fact to care about the aim of getting it right and that they are normatively committed to this aim insofar as they are valuers engaged in the task of practical reasoning. Getting it right, I take it, is an objective matter. It follows that if people are engaged in *that* project, then they can see themselves as failing in some important sense even if they impeccably exercise their capacities for responding to their evidence. These are significant considerations in favor of the fact-relative interpretation.

Still, it is consistent with thinking the fact-relative notion is indispensable (even central) to our thinking about autonomy that the evidence-relative notion has a role to play. As suggested, the evidence-relative notion is clearly needed to make sense of autonomous *beliefs*. However, it may also be relevant to practical reasons. Specifically, it may play some role in funding partial or

indexed autonomy judgments which track agents' unlucky actions. Connie and King George are failing badly relative to an aim that is central to their own self-conception as practical agents: tracking reasons they genuinely have. So they do not enjoy *full* autonomy. But their failure, as it turns out, is all just through bad luck: had circumstances been a bit different, they would be succeeding quite admirably at tracking reasons they have. Moreover, a perfect internal simulacrum of each agent placed in circumstances where the evidence is not misleading, would be perfectly autonomous. Can it really be that each agent's unlucky double is entirely bereft of autonomy? There is no need to say this. It seems quite plausible to say that Connie and King George enjoy, not full, but at least partial autonomy.

In sum, there is reason to hang on to both notions of reasons-responsiveness, one fact- and the other evidence-relative. The reason-first approach supplies a flexible framework to do just this. Note that the sort of tension we have uncovered between more and less objective interpretations of reasons-responsiveness is an implication of the reason-first approach as I have been developing it. For I have stressed the role of beliefs in autonomous agency in two distinct ways: (a) as attitudes with respect to which agents can be autonomous, and (b) as constitutive ingredients of an agent's overall autonomy. These roles can conflict. On the one hand, *holding a belief autonomously* is a matter of an agent's evidence. On the other hand, *being autonomous* (at any rate, being fully autonomous) is not. Put differently, agents are fully autonomous in virtue of their beliefs being correct, but they believe autonomously in virtue of their beliefs being sensitive to evidence that they are correct. It is a straightforward consequence of this view that it is possible for agents to hold beliefs autonomously which impair their overall autonomy.

Capacity vs. fulfillment

The core formula makes autonomy a function of rational control, which it understands in terms of responsiveness to reasons. But is *mere* capacity enough or must there also be some success in realizing this capacity, some actualization of one's rational potential?

The correct account of reasons-responsiveness for the *moral* domain is evidently capacitarian. To be morally responsible it suffices that one has relevant capacities; one need not also exercise these capacities well. The whole point of blame for negative outcomes is that one can be responsible for failing to do what one is supposed to do. When we shift the focus to personal autonomy, by contrast, matters are more complicated. Some normative capacity accounts of personal autonomy appear to be purely capacitarian (Bublitz and Merkel 2009, Kauppinen 2011), while others suggest agents must enjoy some realization of rational capacity (McDowell 2010, Sher 1997). Which is correct? I think these divergent suggestions reflect underlying tensions in our thinking about autonomy. In my view, it would be an unhelpful simplification to choose one of these alternatives and jettison the other. We do better, I think, by attempting to hang on to, and make sense of, both together. The trick is to recognize the different roles played by these ideas in our thinking about autonomy.

Consider Randy. Randy is an adult in his 30s who plays video games all day in spite of the fact that he judges that he ought to do something more valuable with his time. Randy has abilities to appreciate what it would be good to do, has the ability to do those things, and does not suffer from any dearth of valuable options. Yet he stays on the couch. Is Randy autonomous?

In one obvious sense, the answer is plausibly yes. His autonomy competencies are intact and there are no outside influences tampering with his capacities for choice. In this sense, at least, he is also exercising his capacities for autonomy: he is choosing to act in a certain way with his

capacities fully functioning. In another sense, however, the answer is plausibly no. Randy is not exercising his capacities *well*. We feel no positive esteem for him and may even feel contempt for him because of how he is choosing to use (waste!) his capacities. Randy seems to be falling short of a valuable agency ideal.

In chapter 2, I alluded to the possibility that our thinking about autonomy commits us to two distinct autonomy concepts: a responsibility concept and a virtue concept. We can capture our divergent judgments about Randy with the help of these distinct ideas. Randy is a *responsible* agent, but he is not a *virtuous* agent. The fact that he is a responsible agent grounds a strong presumption in favor of letting Randy live with his choice. We might try to persuade him to get off the couch, but at the end of the day, if he chooses to spend his time playing video games, the fact that he is a responsible agent puts much tighter limits on justified interference with his choices than would be acceptable if he were not a responsible agent: the bar of warranted interference gets ratcheted up. I take this to be part of what it is to respect Randy as the kind of being he is. The fact that Randy does not exercise his capacities in valuable ways, however, is grounds for a different set of judgments and attitudes. We do not think that Randy is any sort of model. We do not think he merits esteem for his choices. And we do not think that his subpar choices are particularly worth promoting (even if they must be respected). They don't exemplify anything that we should aspire to ourselves or promote the realization of in others.

Corresponding to these two notions of autonomy are two distinct notions freedom (cf. T.H. Green 1886/1986; Brink 2003: 81). The first is the idea of *having* normative capacity—the ability to detect and pursue norms and values (along with relevant opportunities). This ability can be had even if it lies dormant or is exercised poorly. The second is the idea of realizing normative capacity—of actually successfully tracking relevant norms and values and then successfully

conforming behavior accordingly. This realization is not had if it lies dormant or is exercised poorly. Since the first sort of freedom is what is at issue in constituting us as responsible, we can call it responsibility-entailing freedom. Since the second sort consists in a kind of realization of our rational powers, we can call it perfection-entailing freedom.

I maintain that we do not need to choose between these distinct notions of freedom as interpretations of autonomy. Instead, we do better to recognize the distinct role they play within our thinking. We need the idea of responsibility-entailing freedom to make sense of autonomy as a ground of respect and bar to paternalistic intervention. This is the idea of *de jure* autonomy (cf. Feinberg 1986: 63). *De jure* autonomy is associated with decision-making privileges of certain kinds associated with a bundle of autonomy rights and is closely associated with recognition respect based on that status.[40] Thus, it is agents with the requisite autonomy competencies whose choices qualify for special protections and who are to be accorded the special dignity of being treated like little "sovereigns" within their domain.

We need the idea of perfection-entailing freedom to make sense of autonomy as an agency ideal, something we have reasons to want and pursue, whether for ourselves or others. I argued in the last chapter that it would be odd if parents cared only that their children have adequate options but didn't care about their normative competence over those options. Likewise, it would be odd if parents cared only that their children develop normative capacities and not also about their successful exercise. The same plausibly goes for what we should want for persons more generally, and what we should try to help them achieve to the extent this is in our power.

[40] The qualification, "based on that status," is important. In my view, recognition respect tracks different statuses. So, for example, one might merit recognition respect *as* a human being with a conscious perspective on the world and desires of one's own, but not *as* a responsible agent. Young children are owed the first but not the second kind of recognition respect. This is why I say that responsibility-entailing freedom is *a* ground of respect. I should not be understood as claiming that it is the *only* ground of recognition respect.

Naturally, there are constraints on how the valuable exercise of capacities can be promoted. First, people's rights as choosers set normative limits. Thus, paternalism is often unwarranted. Second, as perfectionists have long argued, there are obvious limits on how one can promote the value inherent in the value itself. The valuable exercise of capacities is an agency achievement presupposing the exercise of the agents' own powers. While one cannot directly bring such achievements about for others, however, one can plausibly promote the value indirectly by helping to secure conditions favorable to its realization (Brink 2019: 14). Third, as I noted in the last chapter, there are complicated questions about who gets to promote whose autonomy. Some philosophers argue the state or other agencies must be neutral as autonomy promoters. Even if that is so, it merely grants an exception to what are otherwise general, standing reasons applicable to all persons. Perhaps acting as an agent of the state, I may not promote the autonomy of my compatriots—at any rate, not under that description. Still, when I leave the office and act as a private citizen, I clearly have reasons to promote the autonomy of my compatriots. More strikingly, so far as generally applicable agent-neutral reasons go, I have reasons to donate money to refugees or school children living in far-away places to promote their living of autonomous lives.

To make sense of the full range of our attitudes and normative judgments, then, we need *both* autonomy concepts—the one associated with responsibility and the other with virtue. While it might be tempting to insist that autonomy is a pure capacity concept, this would be a mistake.[41] To make sense of the full range of our judgments involving autonomy, we need the idea of perfection-entailing freedom as well.

[41] Thanks to David Brink for encouraging me to take the idea of autonomy as a virtue seriously.

4. Elective Control

Here is a further contrast between moral responsibility and personal autonomy. All that moral responsibility seems to require is capacities for tracking and complying with antecedently given norms. There is nothing especially creative or spontaneous about the process. By contrast, personal autonomy seems to call for more. In particular, it seems to call for creativity and spontaneity. Some authors have emphasized this dimension of autonomy by saying that autonomy calls for improvisation (Bagley 2013, Meyers 2004: 39). Autonomy, it seems, is more like jazz than physics—more about creative improvisation than fidelity to determinate and antecedently given facts.

This rings true. Life, after all, does often feel improvised, and most of us will doubt that for every choice we make there is an antecedently determinate fact about which option is correct. Nevertheless, I believe the phenomenon of creativity and spontaneity can be adequately accommodated within the reason-first framework. One way to do so might be to appeal to voluntarism about reasons. According to a hybrid voluntarist model of practical reasons, pioneered by Ruth Chang (1997), some reasons are discovered while others are made through acts of will. If that picture is right, then on the reason-first model, autonomy would be best thought of as requiring responsiveness to different kinds of reasons, some created and others not. Creativity would be accommodated in one's radical capacity to make new reasons by an act of will.

Luckily, this sort of view is unnecessary. Instead of appealing to a controversial meta-normative doctrine, we can appeal to a highly compelling normative doctrine: the idea that normative space is often rationally permissive. The metaphor of "space" is meant to convey the idea of latitude: one has room to chart various courses within permissible bounds. This is an old

idea. Its seeds are found in the Stoic insistence on treating some things as "indifferent" (*adiaphora*) according to reason, even though one might have a preference for one outcome over another, and in early Christian teachings (e.g. on marriage and eating meat) that distinguished between negotiable and non-negotiable religious practices. These seeds, both secular and religious, germinated in medieval and early modern natural law theory with the idea that natural law mandates certain things and leaves others up to individual choice (Tierney 2014). In modern philosophy, the idea of a permissive normative space is perhaps most associated with Kant's distinction between wide and narrow obligation, where the central distinction is about how much discretion or latitude an agent has in fulfilling the duty. Unlike narrow obligation, says Kant, wide obligation "leaves a playroom (*latitudo*) for free choice" (6:390).

The idea is especially compelling in the domain of prudential value or welfare. Most of us probably don't think there is a single best life for any person. Given adequate opportunities and resources, there are plenty of good but incompatible options for most people. When it comes to navigating one's way through the space of options, many different paths are acceptable. Such permissivism does not imply relativism. On the contrary, it is embedded within a view that affirms that some choices are better than others, and that some choices cross the line—being clearly counterproductive to an agent's well-being. The point is just that, within relevant constraints, there is wide berth for individual choice.

How wide that berth is will depend on one's background views about well-being. Hedonism is probably the least permissive doctrine of welfare. Desire-satisfaction, objective list, and perfectionist views are considerably more permissive. Whichever of these one subscribes to, the idea of a permissive normative space does require that one be able to countenance a non-negligible sphere within which people have discretion to make decisions, a sphere in which

rationality does not demand one option over another. The problem of accommodating options is pressing for maximizing consequentialist views (Scheffler 1994). On such views, rationality always demands that one do the very best. If one pairs this with a monist conception of value, rationally permissive choices would presumably turn out to be quite rare indeed. Those who subscribe to such views will therefore not be able to take comfort in my suggestion about how to make room for creative choice within a reason-first view.

In addition to permissivism, I accept the idea that value is plural (cf. Raz 1986, 1997, 1999). I think there are many and diverse goods, that there are many and diverse ways of living a good life, and that there are many and diverse ways of living a good life *because* there are sundry ways of combining values and organizing lives around those values. Moreover, the extent to which different values are commensurable is limited. Hence, the extent to which lives are commensurable is limited. Some lives are objectively very bad; others are objectively very good. But it needn't follow that there is a cardinal ranking scale for all lives.

Suppose then that normative space is frequently permissive. There will often be wide berth for choice between multiple incompatible options which are, in Raz's terminology, equally "rationally eligible" (Raz 1999). This leaves room for creativity and improvisation. If normative space is at many junctures relatively permissive, and if persons must therefore make rationally underdetermined choices, even those who are perfectly responsive to reasons need not be thought of as marching down a narrowly constrained corridor. They would be choosing their route through normative space—choosing one eligible option over another. In that sense, autonomous individuals create themselves by choosing which facts to make true about themselves and their lives. This supplies a ready way to understand the metaphor of self-creation (cf. Ismael 2016). Moreover, given the uniqueness of each of our lives and circumstances, no two individuals will

ever chart precisely the same course through life. This means a certain amount of improvisation or innovation is called for in adapting general normative truths to the particular and special circumstances of one's life. While you can learn from the choices and experiences of others, there is nothing like a playbook which settles your every move.

Let's put these ideas together with rational control. The reason-first model defines rational control in terms of responsiveness to reasons. Yet we are now admitting that reasons sometimes run out. It seems to follow that rational control is possible when normative space is requiring or constraining, but not when it is permissive. When reasons don't settle the matter, something analogous to regal fiat seems to be required: one must simply *decide on* or *elect* a path through the space of alternatives.

Does this mean rationally permissive situations don't call for rational control? It does not. Notice that whether and how rationally permissive options are is not something that is itself up to the agent. Distinguish three situations: (a) those that are rationally demanding (overwhelming reasons in favor of one option), (b) those that are rationally suggestive (slightly more reason in favor of one option over another), and (c) those that are rationally optional (balanced, on a par, or incommensurable options). Agents must be able to discern what sort of situation they are in. In situations of type (a) and (b), rational control means the ability to track the relevant reasons, whether demanding or suggestive. In situations of type (c), rational control means the ability to recognize that it is a permissive situation—and not, for example, to think one is in a permissive situation when in fact one is in a rationally demanding situation. True, when reasons run out, an agent can take any of the eligible courses of action, but such choices are made against the backdrop of rational control—against the background of appropriate sensitivity to the normative landscape. Hence, rational control is operative even in situations of type (c). For convenience, we can

distinguish between *rational control* and *elective control*, where the former is understood in terms of sensitivity to reasons and the latter is the discretionary freedom left over when reasons runs out. Elective control, I suggest, is a genuine kind of *control* insofar as it operates within the bounds of rational control. It will be, to borrow Susan Wolf's (1990) nice phrase, a kind of *freedom within reason*. In short, elective control is not freedom from rational control but a manifestation of freedom within the bounds of rational control.

5. Authenticity

Many current views of autonomy put the idea of authenticity at the center. Very roughly, one is autonomous on such views when one has and acts from authentic attitudes and values—attitudes and values which are really and deeply one's own. The reason-first approach puts the idea of rational control at the center. But this need not mean giving up on authenticity. I want to suggest that there is a positive role for the idea of authenticity to play in the reason-first approach to autonomy.

First, however, note two benefits of the reason-first approach to thinking about authenticity. First, normative capacities help to make authenticity more robust. If one grants that what makes options choice-worthy is the reasons bearing on them, then the sense of ownership becomes thinned out in the absence of appropriate sensitivity to the relevant reasons. Normative capacity is an additional bulwark of agential independence, securing the agent against inner and outer threats to having attitudes and values that count, in an appropriately deep sense, as one's own. Second, against the backdrop of reasons-responsiveness, pressure is decreased to settle on one unique authenticity relation as all-important. Since formal theories eschew substantive normative

capacities, they cannot afford to be so relaxed. The absence of substantive normative capacities puts great pressure on authenticity conditions to pick up the slack and secure a sufficiently robust notion of agential independence. The reason-first approach can afford to take a more relaxed approach to authenticity. There are plausibly a range of important authenticity phenomena: centrality to one's evaluative outlook, coherence with practical identity, emotional attachment, willful commitment, deliberated endorsement, preparedness to answer for oneself, and so on. Theoretical pressure is reduced to single out any of these authenticity phenomena as definitive or criterial of *real* authenticity. Because less rides theoretically on authenticity, there is less pressure to decide just which relation is definitive of the agent's true or deep self.

Now for the role authenticity plays within a reason-first approach. There are at least two such roles. The first is to help identify the deep self as an object of respect. Part of respecting persons is honoring (within appropriate limits) their point of view and wishes flowing from that point of view. This is vital for respecting persons' claims to conscience. Chapter 2 discussed the case of a Jehovah's Witness refusing a blood transfusion. This case dramatizes the point that some weight must be given to a person's wishes flowing from their deep evaluative commitments. Less vexing and more mundane cases are common. For example, suppose you are an executor of a living will that stipulates money should be given to an organization you despise. Suppose you also know that the person whose will it is decided to give to the organization on a whim and without any research; had they actually done their homework, they would have realized that, they too find the organization despicable. Contrast this with a case where the person is well-informed about the organization and where that organization reflects her deeply held and stable values and convictions. Reasons of respect would seem much stronger in this second case. Of course, you may be legally bound to carry out the will's provisions in any case, and there are complications

about what it means to respect a deceased person. But these caveats are not essential to the point that authenticity helps define a person's deep self, and a person's deep self makes a difference to reasons of respect.

The second role played by authenticity is as an aspirational value. Here we might think of the kind of formation of self that is such a central project during adolescence and early adulthood. There are presumably many different and legitimate developmental pathways both within and across different cultures. Nevertheless, there is something sad and stunted about a young person who fails to develop an independent sense of self. There is, in short, some importance to the emergence and cultivation of distinctive selves. It may be that the crystallization of an independent vantage point on the world is valuable for its own sake. But *merely* being "one's own" person is hardly something worth striving for unless one's self is sensitive to, and organized around, genuine values. The reason-first approach can thus recognize the importance of forging an agential perspective on the world, while preserving the idea that having such a self would actually be of much greater value if combined with orientation toward, and successful pursuit of, genuine value.

6. Independence of Mind

The reason-first approach gives us resources for thinking about what it is to have an independent mind as well. Formal views may seem to be better resourced here, since whether or not one has an *independent* mind seems to be separate from whether one has a mind that is appropriately *responsive to reasons*. A confident person trusts her own judgments. This is surely not the same as actually being responsive to reasons. As I shall explain in a moment, however,

being responsive to reasons does help to secure independence of mind. Since formal views eschew responsiveness to reasons, they cannot avail themselves of this additional resource.

Recall the discussion about reasons-responsiveness and objectivity. We can distinguish more and less objective senses of responsiveness to reasons:

- Responsiveness to reasons *as we see them*
- Responsiveness to reasons *as indicated by our evidence*
- Responsiveness to reasons *as they are*

I argued above that reasons-responsiveness should be understood as evidence-relative for the attitude of belief but fact-relative for practical reasons in general. I now want to add that we can make use of the least objective sense of reasons-responsiveness as the first rung on a ladder of ascending mental independence, where responsiveness to reasons as they are is the highest-grade kind of mental independence.

Being disposed to operate on reasons *as one sees them* constitutes a first level of mental independence. On at least one reading, the problem with the deferential housewife is that she does not operate on reasons as she sees them—all of this is outsourced to her husband. Benson (2005) suggests that her lack of self-respect causes her not to trust her own judgments. Westlund (2003, 2009) claims that the deferential housewife's systematic outsourcing of her judgments is inconsistent with a disposition to answer for herself—to cite reasons as she sees them, when challenged to do so.

While being disposed to operate on reasons *as one sees them* does afford some minimal level of mental independence, it is important to note just how minimal it is. Westlund admits, for example, that as long as the wife defers to her husband because that accords with her values and is how she sees the reasons she has, then she is perfectly autonomous. Thus, merely operating on

reasons as one see them clearly does not secure a very robust sort of mental independence. It is compatible with being brainwashed or duped, as also with having a seriously deferential cast of mind (excluding only its most radical form in which one cannot even cite reasons for one's deference).

Being disposed to operate on reasons *as indicated by one's evidence* constitutes a second level of mental independence. If the wife operates not only on reasons as she sees them but on reasons indicated by her evidence, then she will presumably not value deference to her husband as such, since such deference will not (in normal circumstances) be indicated by her evidence. Mental independence of this kind secures agents more robustly. It will rule out more cases of brainwashing and duping and deference than merely operating on reasons as one sees them. Still, it doesn't rule out all such cases. Suppose the wife's evidence is radically misleading. Paul Benson (1994) discusses a case of "gaslighting." Taking inspiration from the 1944 film, *Gaslight*, by Ingrid Bergman, Benson (1994: 656) imagines a woman who "falls into a state of helplessness and disorientation as a result of a profound change in her view of herself." Her husband is a physician who, on the basis of the accepted science of the day, regards emotionally excitable women with active imaginations as mentally ill. "The protagonist has the suspect traits, her husband makes the standard diagnosis, and the 'hysterical wife' ends up isolated, feeling rather crazy" (656). It is key to the way Benson sets up the story that the wife has good reason to believe on the basis of the evidence she has that she is hysterical. The problem is not that she lacks the disposition to respond to reasons as indicated by her evidence; it is that her evidence is deeply misleading. While being sensitive to one's evidence presumably makes it *harder* to be manipulated, it doesn't make it impossible.

Finally, being disposed to operate on reasons *as they are* constitutes a third level of mental independence. This level secures the most robust kind of mental independence. For it rules out cases of the sort imagined by Benson. To use terminology introduced above, if one is fact-relative reasons-responsive and not just evidence-relative reasons-responsive, one's beliefs will have to be substantially correct. Because she has been misled and her beliefs are systematically in error, the gaslighted woman is not fact-relative responsive to the reasons she has, though she may be responding well to the evidence she has. Being fact-relative reasons-responsive adds further protection against manipulation by others.

There is a further sense of mental independence which, though related to these reasons-responsiveness notions, is not the same. We are all familiar with the phenomenon of subtly modulating one's opinions under peer pressure and, more generally, with yielding under pressure to conform. Having an independent mind can also mean having a certain resilience against pressure to conform, give up, change one's mind, and the like. Independence of mind, in this sense, is a disposition of character. I think of it as a "formal" or "structural" virtue, like courage, loyalty, principledness, circumspectness, reflectiveness. What these virtues have in common is that, unlike substantive virtues like justice or benevolence or wisdom, they are not good or valuable in themselves and may even be bad. Because of our human frailties, they are *typically* good traits to have, and they can play an important supporting role in making us substantively virtuous. But they don't suffice for this. One may be courageous and loyal and principled in the service of very bad causes. Their value is thus largely instrumental. Having an independent mind in this sense is a formal or structural virtue because it can aid the individual in attaining reasons-responsiveness in the three senses just considered, though it need not do so without fail. If one exhibits resistance to peer pressure, willingness to take risks, resilience in the face of opposition, unwillingness to

conform or fall into line for the sake of the associated rewards, and so forth, one will tend at the very least to operate on reasons *as one sees them*. But given our typical frailties as humans, possessing this virtue may also make us a bit more responsive to our evidence and, one may hope, responsive to reasons given by the facts.

7. Autonomy's Milieu

Epictetus (2018), the famous slave-turned-Stoic-philosopher, is a good exemplar of a certain tradition of thinking about freedom. He writes as if freedom were an entirely internal affair, available to any agent in any circumstances—even to the slave. The resulting picture is starkly internalist. Governance by Reason within the individual soul does not depend on external circumstances. One can be self-governed and free, no matter how shoddy the situation, since one can always do whatever it is one has most reason to do in the situation

In emphasizing rational control, the reason-first approach to thinking about autonomy has some affinity with this older rationalist tradition of thought. However, there is little reason to hang on to the austere vision of freedom championed by Epictetus and other Stoics. On the contrary, the best version of a reason-first approach will make autonomy depend, in at least two ways, on factors outside the agent.

First, an agent's rational abilities are not cross-situationally invariant. As a rich body of literature in social psychology demonstrates, human behavior and rational competence are to a significant extent environmentally conditioned.[42] A number of responsibility theorists have

[42] This context-dependence of human thought and behavior is apparent in its sensitivity to a host of contextual factors, often quite arbitrary (Wilson 2004, Ross & Nisbett 1991/2011). A number of classic studies show what a profound effect on moral and prudential performance such context-dependence can have (Darley & Batson 1973, Isen & Levin 1972, Kahneman 2011, Milgram 1969/2009, Mischel et al. 1989, Thaler & Sunstein 2009, Zimbardo et al. 1973).

concluded, rightly I think, that the very capacity for reasons-responsiveness is sensitive to ecological factors outside the individual's direct control. An agent's psychological infrastructure for reasons-responsiveness—her normative competence—is relatively cross-situationally stable, but her ability to actually deploy that competence in a given situation—her situation-indexed capacity—is more contextually variable (Brink 2013, Brink & Nelkin 2013, Nelkin 2005, Vargas 2013b, 2017, 2018). In short, reasons-responsiveness is, at least in part, dependent on factors *outside the head*.

Second, if autonomy is to be a meaningful personal and social ideal it must go together with a broader social vision that includes both constraints and options. What I want to emphasize here is that these external conditions aren't ad hoc additions; they grow quite naturally from the account I have been developing. I have argued that normative capacity is a source of negative autonomy rights and ground of recognition respect. These are constraints on how we may treat others.

Following Raz (1986), I have also argued that valuable options matter for autonomy. To cite one real-world example, economists Betts and Collier (2017, ch. 6) write about the lack of opportunities of most refugees living in camps as a hindrance to their autonomy. These claims are surely plausible, and they are supported by many who write on autonomy.[43] On the reason-first approach, there is a principled rationale for accepting the importance of options. As I have argued, autonomy makes self-creation (in a limited sense) possible; it also constitutes persons as responsible; and it is the sort of thing we have reasons not only to respect but also to promote, i.e.,

[43] See, for example, Kauppinen (2011: 284ff.), Mackenzie (2014: 28), Mackenzie and Stoljar (2000: 22, 26), Oshana (1998: 94), Terlazzo (2016). A number of authors, e.g. Friedman (2004: 159-161), Kauppinen (2011: 281), Stoljar (2014), Terlazzo (2016), have noted that adaptive preferences are suspect from the point of view of autonomy. If autonomy requires valuable options this would explain why there is plausibly something suspect about adaptive preferences from the point of view of autonomy.

it is a *valuable* ideal. Putting these ideas together yields a plausible defense of the importance of having adequately valuable options.

Part of what a credible and attractive self-creation ideal must mean is that people are responsible in a suitably rich and meaningful sense for their lives, and that means that their lives cannot be forced upon them by stunted options. Like a competent artist, with imagination, skill, and a realistic sense of her medium, the autonomous person makes something (a life) which is, in some sense, the product of her activity. This requires internal competencies, as I have argued, but it also requires external resources—a sufficiently good medium and tolerably supportive conditions to carry out her artistic labors.[44] To be able to exercise and realize their normative capacities, people must have sufficiently valuable "material" to work with. The importance of options is therefore supported by twin considerations of responsibility and value. The person who chooses from an extremely limited menu is less responsible for the shape of her life than an equally capable person choosing from a richer menu. Moreover, the better the menu, the greater the opportunity for living valuable lives. When we reflect on what we have reason to promote, valuable options are part of the package.

Of course, people who suffer from a lack of options need not thereby suffer from a *complete* lack of autonomy. We can easily imagine two refugees, faced with the same meager resources and opportunities, one of whom lives more autonomously than the other. Autonomy, as we have noted, is a scalar property. Thus, a dearth of options need not mean complete lack of autonomy. Moreover, as we have also noted, autonomy is multi-dimensional. One can be fully autonomous in one dimension without being fully autonomous in another. Some philosophers might think it the mark

[44] Wall (1998: 142) opts for a subjective interpretation of the adequacy of options: they must good enough for people to live the lives *they want to live*. The rationale I have given speaks in favor of a more objective adequacy criterion, though assuming there is rich variety among objectively valuable options, it will accommodate a wide range of reasonable patterns of subjective preference.

of an adequate account of autonomy to deliver all-things-considered judgments about where any individual ranks on a scale of autonomy. I don't share that conviction. Given that autonomy is multi-dimensional, there is no reason to think comparisons across dimensions must always be meaningful. Compare Randy, the well-resourced video game player who lays on the couch all day, with Hassan the refugee, who has few resources but applies himself creatively and vigorously to make a life from what he has available. Who is more autonomous? The question cannot be meaningfully answered. Each agent enjoys a good the other lacks: Randy enjoys a kind of freedom Hassan lacks; Hassan enjoys a kind agential excellence Randy lacks. Each agent therefore enjoys more autonomy than the other *in one dimension*.

In principle, one can imagine a device whereby the different dimensions are summed together to yield an overall verdict. This might especially be so if one thought one or another dimension of autonomy has greater value than the others. One might then add greater weight to this dimension in calculating the total. Assuming each dimension was assigned some weight, one could then compute a sum total for all agents on the same scale. In practice, this sort of summing device is not likely to be illuminating. First, there is no answer to the question, which dimension of autonomy is most important? Our autonomy-talk tracks different things. So the answer depends on context. For example, are we talking about agency ideals worthy of promotion? Or agency capacities grounding reasons of respect? Second, if the dimensions are really distinct, then all-things-considered judgments summing across different dimensions obscure important information. Thus, if we say Hassan is more autonomous than Randy or vice versa, we lose valuable information. We retain that information with appropriately qualified or indexed judgments: Hassan is more autonomous than Randy *in respect of X*, but Randy is more autonomous than Hassan *in respect of Y*.

The reason-first approach to thinking about autonomy is thus compatible with affirming the importance of external constraints and options. It is not committed to Epictetus-style freedom which would seek to make autonomy robust against extremely wide variation in circumstance. Quite the opposite. On realistic assumptions about our limits and fragilities as agents, and our deep dependence on material and social conditions, the reason-first approach recommends a picture of autonomy starkly at odds with Epictetus-style freedom.

I suggested in chapter 2 that rejecting formal views of autonomy is consistent with political liberalism. This is not the place to explore the political implications of the reason-first approach in any detail. I merely note the following implications. Personal autonomy needs a milieu within which to flourish. In broad outlines, such a milieu requires a social and political world characterized by (i) strong norms of respect for persons in their capacity as responsible choosers, and by (ii) the presence of a rich array of valuable options. It will be a world in which people have adequate opportunities to live very good lives but not a world in which they are forced to live those lives. Concretely, this suggests a social and political environment characterized by fairly robust anti-paternalist norms and adequate material resourcing. The reason-first approach would also put a premium on education. Since education empowers persons to develop their normative capacities, this would be a further vital resource in an autonomy-conducive milieu.

What more precise policy implications might follow, I leave open. For example, would a reason-first approach favor nudging (cf. Thaler & Sunstein 2009)? That may depend. On the one hand, the ecological perspective on reasons-responsiveness I have plugged for would appear to speak in favor of nudging. On the other hand, there are obvious constraints given by the importance of respecting autonomy and promoting its optimal exercise. My suspicion is that a piecemeal

approach makes most sense. Some nudging will be respectful of autonomy; some will not. Some nudging will promote autonomy; some will undermine it.

As I have stressed, the reason-first approach to thinking about personal autonomy is compatible with a variety of positions about politics. The social vision it supports as most hospitable to autonomy fits very naturally with the kind of perfectionist liberalism championed by John Stuart Mill and T.H. Green, which derives quite comprehensive liberal principles and policies from a conception of persons as normative agents whose capacities must be protected and promoted (see Brink 2007, 2013). But one might take a less ambitious approach, preferring to see liberalism as a complicated set of doctrines, not all of which ground in the same core vision of personhood and normative agency. In that case, the protection and promotion of normative capacities would be just one part of one's political vision, and maybe a small one at that. Finally, one might accept a reason-first approach to thinking about autonomy but insist that it is no legitimate aim of the state to promote the autonomy of its citizens. In that case, while what I have said about autonomy's milieu would not be permitted to have political implications, it would presumably still have private implications. Persons would still have reasons to protect and promote autonomy-friendly circumstances for themselves, for their loved ones, and for strangers near and far.

Conclusion

A reason-first approach to thinking about autonomy makes certain assumptions. In particular, it assumes that there are genuine reasons and values, that practical reason has as its aim the tracking of these reasons and values, and that responsibility—or rather, one kind of

responsibility of which persons are capable—is to be understood in terms of abilities for responding to these reasons and values. As I attempted to show, these assumptions are both credible and reasonably ecumenical. Though not uncontroversial, they are highly plausible and can be shared by a broad range of views.

One might build on these basic assumptions in more ways than one. My emphasis has been on the idea of rational control, understood in terms of reasons-responsiveness. I argued that this idea comes with great power and flexibility: it licenses scalar and threshold judgments, allows us to make person-level and sub-person level judgments, helps us think both about what it would mean to be autonomous in respect of actions and attitudes, can accommodate different senses of objectivity, and holds the key to the distinction between the responsibility and virtue concepts of autonomy. Moreover, I argued that this framework leaves room for personal creativity and spontaneity—or at any rate, it does if we are prepared to accept something that seems quite plausible, namely that normative space is often rationally permissive. Finally, I argued that putting rational control at the center of our view of autonomy allows us to thread a variety of distinct autonomy concepts (and their associated values) together into an attractive package. It helps us distinguish and hang on to autonomy as responsibility-entailing and autonomy as perfection-entailing freedom, the first associated with recognition respect tracking responsible agency capacities and with negative autonomy rights, the second associated with performance respect and an agency ideal worthy of promotion. Moreover, the reason-first perspective gives us resources for thinking about authenticity, inner independence, and external independence. It makes good on authenticity as a compelling agency ideal while relaxing the pressure to settle on any single criterion of authenticity; it supplies plausible ingredients for thinking about different kinds and degrees of intellectual independence; and it yields a natural way to think about what sort of external

environment would be needed for autonomy in terms of the idea of conditions hospitable to the exercise of rational capacities.

The reason-first view thus has considerable explanatory power. It does not recover ordinary autonomy-talk in any conceptually simple way. Instead, it gives us something else. From a patchwork of concepts and normative judgments associated with our thinking about autonomy, the reason-first view reconstructs an interpretation of autonomy that seeks to square maximal fidelity to complexity with maximal theoretical simplicity. It supplies us with a basic picture of responsibility and normative agency, and with a rough-and-ready formula for putting that picture to work across a wide range of phenomena and normative concerns. To some readers, this way of thinking about autonomy may smack of intellectual imperialism. The view, they will worry, seeks to cover too much ground and draw too many strands together. Isn't it better to focus on each thing in its particularity? I agree that theoretical overreach is a worry. In the end, if the reason-first view as I have been developing it here overreaches, it will have to be either pruned or rejected. My goal has been to sketch the positive case for the view by giving a tour of its rich resources and possibilities. If I am right that it offers a powerful and flexible framework for thinking about autonomy, that certainly counts in its favor.

Chapter 4

The Activity of Self-Governance

The last chapter developed a substantive alternative to formal views of autonomy in terms of the idea of rational control. This chapter explores what I call the activity model of self-governance whereby agents manage themselves in the service of valued goals. Understood in this way, self-governance is similar to what Robert Adams (2006) calls a structural virtue. While it is not itself a substantive notion, self-governance in this sense plays an important role in realizing and sustaining substantively rational agency and is worth exploring for that reason. Most of us do not enjoy perfect responsiveness to reasons all the time. Instead, we must work at discerning what reasons we have and exert some effort in trying to align our lives with our appreciation of reasons. This requires active self-management.

To forestall confusion, it is worth registering a caveat at the outset. I spell out the activity model of self-governance in terms of top-down regulatory control. Since I interpret self-governance in structural terms, however, the operative notion of control in this and the last chapter differ. The notion of control invoked in this chapter is non-normative: it has to do with an agent's ability to bring about a desired goal-state, whatever the goals happens to be. The notion of control invoked in the last chapter is normative: it has to do with an agent's ability to appreciate and conform to relevant norms. We need each notion for a different purpose. I argued in chapters 2 and 3 that our thinking about autonomy commits us to the relevance of normative capacity and its realization. For these purposes, the operative notion of control has to be normative. But in this chapter, I turn to active self-management as a formal feature of agents. For this purpose, a less

demanding notion of control is needed, one that does not require, though it is entailed by and can play a role in, responsiveness to reasons.

The chapter proceeds as follows. Section 1 begins by clarifying the notion of self-governance that is my focus in this chapter. Section 2 gives examples of the phenomenon. Section 3 spells out the activity model of self-governance in terms of the idea of top-down regulatory control. Sections 4 and 5 enrich the picture by considering aspects of self-governance that transform this kind of control into something that, while deeply continuous with the rest of the animal world, transforms it into something characteristic of distinctively human agency: conscious reflection and normatively articulated ends. Sections 6 and 7 discuss the relationship between self-governance interpreted in the structural sense and the substantive goods of responsiveness to reasons and substantive virtue.

1. Three Notions of Self-Governance

The language of self-governance can be used in different ways. Consider three notions of self-governance:

1. *Expressive Authenticity Model*: People govern themselves when they act from authentic motives, that is, from motives which are in a suitably deep sense their own.

2. *Rule-of-Reason Model*: People govern themselves when they act rationally or in accordance with reason.

3. *Activity Model*: People govern themselves when they engage in activity that has as its aim the regulation of behavior.

The first idea is expressed by Laura Ekstrom (2005: 155) when she writes, "One's action is self-governed when it is directed by the true self." This way of using the language of self-governance has gained widespread traction in the moral psychology literature since Frankfurt (1971). It is perhaps most associated with the work of Michael Bratman (2000, 2003, 2004, 2009), who argues that the privileged set of attitudes constituting an agent's standpoint are closely connected to an agent's self-governing policies.

The second idea originates with a picture developed by Plato in the *Republic*, according to which the element of a person's psyche that ought to govern is Reason (411c, 431a-e, 441e, 442c). Reason is the soul's *proper* authority or ruler: it ought to be commanding, the passions obeying. The idea has a long afterlife in the history of philosophy. For example, writing in the early 18th century, Joseph Butler (1726/1983) claims that the "principle of conscience or reflection" (29, 30, 37), by which persons "approve and disapprove their own actions," is the faculty that ought to call the shots. This "highest principle" ought to "preside over and govern all the rest" (34); "it was placed within to be our proper governor, to direct all under principles, passions, and motives of action" (40). The principle of conscience or reflection is said to enjoy a natural "prerogative," "supremacy," and "superintendancy" (38, 40). Butler even recycles Plato's city/soul analogy: "And as in civil government the constitution is broken in upon and violated by power and strength prevailing over authority, so the constitution of man is broken in upon and violated by the lower faculties or principles within prevailing over that which is in its nature supreme over them all" (41). Rehabilitating the tradition for a contemporary audience, Christine Korsgaard (2009: 131-

158) explicitly defends the Platonic model, arguing that an agent's true constitution puts reason in charge, giving it the authority to govern. When people fail to act from reason, they are not identified with their proper constitution as agents, so they fail at governing themselves. More radically, on Korsgaard's reading, this turns out to mean that when people are not identified with their proper constitution, they either do not act or else act defectively (152, 159-176).

The third idea is exemplified by the classic story of Ulysses and the sirens. Ulysses desires to hear the siren song, but not to recklessly endanger his life or the life of his men. In order to safely indulge, Ulysses has himself tied to the ship's mast and orders his men to put wax in their ears and to disregard any orders or protests from him until the ship has sailed past a certain point. Ulysses displays self-governance. In this case, the agent makes creative use of circumstantial conditions to deal with anticipated temptation and safeguard appropriate behavioral outcomes.

In the philosophy and psychology literature, activity of this kind often goes by other names, like self-regulation, self-control, and self-management. Nevertheless, the language of governing seems apt. For example, Al Mele (1995: 139) writes, "Provided that a desire is not irresistible, one may be in an excellent position to govern its influence—or to prevent it from influencing one's overt behavior at all." To "govern," in this sense, agents must *do* various things to bring it about that they act (or don't act) in certain ways. They might need to directly resist urges and impulses through effortful acts of will. Or they might need to do a variety of less direct things: reflect and deliberate, make earlier plans and commitments to guide later behavior, change the environment, change the incentive structure of future action, distract themselves, re-imagine their situation, and so on. Either way, governing in this sense involves active self-intervention.

While Plato is associated with the rule-of-Reason model, it is worth noting that he also has a notion of self-governance closer to the activity model. Indeed, the language of *governing* in the

Republic is primarily formal or procedural: it conveys the idea of top-down control. The Greek term for rule is *arche*. *Arche* conveys the idea of superior and inferior, of one who commands and another who obeys (Long 2015: 129). Plato describes rulers in functional terms, i.e., by what they do: they command and make laws (458c, 465a, 502b). Hence, it is perfectly intelligible for Plato to say that a state may be ruled unjustly (362b). Likewise, he can say that a ship may be governed in an unruly fashion (488b) and that souls may rule themselves badly (353e). Rule or governance is not *ipso facto* good rule or governance; it can be appropriate or inappropriate (444b), according to or contrary to nature: "Then, isn't to produce justice to establish the parts of the soul in a natural relation of control, one by another, while to produce injustice is to establish a relation of ruling and being ruled contrary to nature?" (444d). As these passages suggest, *arche* does not itself convey the idea of right or good rule. Plato reserves various virtue terms—wisdom, moderation, justice—for the achievement wherein governing in fact aligns with Reason (cf. 430e, 442c). In short, Plato thinks of the activity of ruling or governing as defined primarily in functional terms, whereas he thinks of ruling or governing well in primarily substantive terms.

The first two notions of self-governance convey an ideal. According to the *expressive authenticity model*, people govern themselves when they act from motives which are, in some significant sense, their own. According to the *rule-of-Reason model*, people govern themselves when Reason is in charge. On either model, many actions fail to meet the condition specified by the ideal: only a privileged subset of actions will be such that they reflect an agent's authentic self or are in accordance with Reason. Viewed in this way, self-governance is (i) an achievement (that is, it is not characteristic of a great deal of human activity), and (ii) something which is supposed to be inherently desirable. Contrast this with the third notion of self-governance. According to *the activity model*, people govern themselves when they engage in characteristic sorts of activity,

including planning, control of appetites, regulation of affect and attitude, and habit management. Self-governance in this sense is (i) not an achievement (hence, it is characteristic of a very wide swath of human behavior), and (ii) not inherently desirable (though it may have instrumental value).

My focus in this chapter will be on the activity model of self-governance. This is not because self-governance in the sense conveyed by that model has great independent value. On the contrary, as I see it, self-governance in this sense has no independent value: its value is entirely instrumental—valuable because, and only insofar as, it helps realize other goods which have non-instrumental value. But this instrumental value also makes it worth exploring in the context of the present dissertation, since it has importance for realizing the goods associated with autonomy. Granting that authenticity and alignment with Reason are each desirable in their own way, becoming authentic or aligning one's life with Reason is not at all a given; it often requires some work on the agent's part. For one thing, it requires forging an authentic self and discerning what it is one has reason to do. For another, it requires implementing these understandings against the push and pull of circumstance, the recalcitrance of motivation, the inertia of habit, and so on. People don't automatically have authentic selves and know what they have reason to do; nor do their lives automatically line up with their authentic preferences or with how they have reason to live. Realistically, then, in order to attain self-governance in an ideal sense, people must actively engage themselves and their circumstances. The activity model is therefore worth exploring as a (partial) implementation mechanism for the achievement of personal autonomy. It holds the key to understanding concretely how people can exercise and achieve autonomy.

To be clear, what I am here calling self-governance might be called something else, like self-management or self-stewardship or active self-intervention. Readers who prefer to reserve the

language of self-governance for one of the ideal models can think of what I am investigating in this chapter under one of these alternative labels.

2. Adumbrating the Phenomenon

The phenomenon I have in mind is nicely illustrated by Walter Mischel's classic delay-of-gratification studies. In these studies, four-year-old children were presented with the choice between a tempting present option, which could be consumed earlier, and a more highly valued distal option, which could only be consumed later. Mischel and colleagues (1989, 2014) observed that children who successfully waited for the later reward often used self-distraction tactics, like looking away, covering their eyes, or signing songs. The experimenters varied many different experimental parameters to see which ones tended to either facilitate or undermine delay of gratification. Confirming their informal observations, when children were prompted to think distracting thoughts, they showed enhanced self-control and waited significantly longer. An even more effective self-control strategy involved cognitive construal of the situation. Prompting children to focus on the "hot," arousing features of the stimulus consistently reduced wait-time, whereas prompting children to focus on the "cold," abstract or informational qualities of the stimulus consistently lengthened the time they managed to wait.

The young children in Mischel's studies exhibit varying levels of executive self-regulation in the service of a valued goal. Resisting contrary impulses and staying the course requires effortful control. On one model, effortful control requires willpower, which is a bit like a muscle—a limited and energetically expensive resource that can become exhausted or replenished (Baumeister, Vohs, and Tice 2007). However, the use of attention-shifting and cognitive construal by the children

suggests that more is going on than raw exertion of willpower. These cognitive control strategies enable the children to gain some mastery over their impulses through *indirect* means. The children do not simply face their urges head, quelling or overpowering those urges through a direct act of will. Instead, they engage in techniques of self-manipulation by distracting themselves from the arousing situation or mentally focusing on its less arousing features.

Like children, adults engage in active self-regulation, though they do so in a wider set of contexts and with additional cognitive and volitional resources that considerably expand the arsenal of self-management techniques. The basic self-regulatory strategies for the control of appetites and affect are the same, including situational alteration (removing a stimulus or triggering condition), attentional redirection (ignoring the source of arousal), and cognitive re-appraisal (deploying the imagination, shifting one's perspective) (Gross 2014, 2016). Cognitive re-appraisal appears to be a particularly powerful tool for emotion-modulation (Kross & Ayduk 2011). These resources can be put to use in cognitive behavioral therapy (Hofman et al. 2012) and in what Timothy Wilson (2015) calls "story editing," learning to reimagine one's situation in ways that can lead to positive behavioral change.

An important set of tools in the self-management arsenal consist in manipulating one's body or environment. For example, to deal effectively with one's anger, one might need to leave the room. To get anxiety under control, one might need to meditate. To sort through a relationship or make a career choice, one might need to spend some time journaling. These techniques work through altering physical location, manipulating posture and breathing, and using an external writing device, respectively. Self-manipulation of this kind is in fact ubiquitous and mundane. Think of the way most of us consume substances and carefully curate our environments to deal with states we experience as mostly "passive" like moods, emotions, and desires. To get motivated,

one might drink a cup of coffee and change the background music to a more up-beat tune; to feel better, one might settle down with a glass of wine and distract oneself with a movie.

A large portion of self-managing activity consists in the deployment of goal-setting and planning (cf. Bratman 1999, 2018). What I want to draw attention to here is the fact that effective planning agency is suffused with *indirect* control strategies. The story of Ulysses, mentioned above, dramatizes the point. But as Thomas Schelling (1978: 290) notes, the phenomenon is utterly quotidian:

> Many of us have little tricks we play on ourselves to make us do the things we ought to do or to keep us from the things we ought to foreswear. Sometimes we put things out of reach for the moment of temptation, sometimes we promise ourselves small rewards, and sometimes we surrender authority to a trustworthy friend who will police our calories or our cigarettes. We place the alarm clock across the room so we cannot turn it off without getting out of bed. People who are chronically late set their watches a few minutes ahead to deceive themselves.

Consider one final self-management technique: altering habits. Habituation plays an important role in human action, allowing behavior to become routinized and automatic, thereby relieving executive decision-making and freeing cognitive space. Since habits use associative learning mechanisms which pair cuing contexts to performance, an important locus of intervention is to seek to alter the cuing context (Neal et al. 2006). To break a bad habit, it helps to avoid situations that tend to trigger the behavior. As Aristotle recognized long ago, habits are vital sites of self-sculpting. Since we do not enjoy direct control over our habits, trying to change our habits is another example of an indirect control-strategy.

3. Top-down Regulatory Control

The activity model of self-governance holds that agents are active self-interveners: they *do* various things to regulate their conduct (cf. Roskies 2012, Vierkant 2013). At the heart of this model is the idea of top-down regulatory control. To get a better handle on top-down self-regulatory control, it will be useful to make three distinctions.

Directness. The examples I have given—Mischel's marshmallow test, cognitive re-construal of a painful emotion, meditation, limiting one's future options, changing locations—involve interventions on motivational, representational, and affective states, and manipulation of body or environment. It is important to recognize the continuity of such self-intervening activity with much ordinary and uncomplicated activity. The agent's ability to self-intervene depends on the agent *doing* various things. Hence, self-intervening activity depends on and recruits more basic capacities for intentional action. The simple activity of making a cup of coffee manifests an agent's basic control capacities. If the agent decides to try to kick her coffee addiction, the more complicated and indirect forms of self-manipulation she engages in to try to kick the addiction will be continuous with, and depend on, the elemental control capacities she displays in the simple act of making coffee. In order to kick the addiction, the agent will need to *do* various (sets of) simple things: make an "implementation intention" (cf. Gollwitzer 1999) not to pick up coffee at the grocery store, engage in a pre-commitment strategy (cf. Elster 1979, 2000) by ridding herself of her coffee maker, imagine the coffee as a toxic substance or distract herself when she is feeling cravings (cf. Mischel 2014), call a friend for support, and so on.

Complex self-interventions occur in cases where agents lack direct control. Valeria wants to practice the violin more often but finds that she frequently lacks motivation. To deal with her lack of motivation, she decides to engage in activity over which she has more direct control. To drum up motivation, she watches YouTube videos of Itzhak Perlman to remind herself of why she

has decided to learn violin. To combat temptation, she practices in a room other than the one where the videogames are, makes the rule that she may only play an hour of videogames for every half hour of violin practice, and tells all this to her accountability buddy. The key here is that some things are more directly within Valeria's control than others: she uses actions that are more directly in her control to influence behavior that is less directly in her control. She does not have direct control over her motivation, but she does have direct control over whether to watch a YouTube video or where to practice the violin.

D*egree of control*. The degree of control an agent enjoys is a measure of what it is in her power to do. At the upper limit, we might say you have complete control if it is, in some sense, entirely up to you whether you whether you X, and at the lower limit we might say you have no control if the occurrence of X is not up to you at all (cf. Mele 2017). A more precise characterization of degree of control is available in terms of counterfactual success at reaching the goal state which is the object of controlled activity (cf. Shepherd 2014). In Joshua Shepherd's (2014) example, the accomplished dart-thrower and the novice may both hit bull's-eye, but the accomplished dart thrower has greater control than the novice over the relevant outcome. What this difference between the players comes to has to be cashed out in terms of the range of counterfactuals in which they each would hit the target. Each player may equally desire to hit bull's-eye, yet one of them will more consistently hit the target across wider variations in circumstance.

Like knowledge and virtue, control is a modal notion, which must hold across some range of circumstances and rule out merely getting lucky. If you just happen at 3:15p to glance at a broken clock that is stuck at 3:15p, and you conclude on this basis that it is 3:15p, you do not know that it is 3:15p. If you are moved in a moment of compassion to help an elderly woman across the

street but this is because you just received good career news and are feeling a bit more upbeat than usual, this does not count as a virtue. Similarly, if you are driving on ice and happen not to skid, this does not mean you are in (much) control of the car.

Degree and directness of control are cross-cutting distinctions. While they will often go together, they need not. The highly experienced meditator who is able to reach a state of emotional calm within one minute of assuming her pose and regulating her breath and who can reach such a state reliably across many different situations, enjoys a high level of indirect control. The drunk person who has difficulty speaking enjoys a low level of direct control over her speech.

Top down vs. bottom up. Top-down regulatory control can be characterized in terms of hierarchically structured information-processing. Of course, information flows the other way as well. Any goal-directed action requires sophisticated forms of self-monitoring and self-adjustment, feeding information about the state of the agent vis-à-vis the goal, and about the state of the goal (e.g., about its continued availability and desirability), back into processing, so that processing is continuously updated in feedback loops. Nevertheless, in top-down regulatory control, higher-level processes regulate lower-level ones. It is natural to describe the relationship between higher and lower-level processes metaphorically, in terms derived from hierarchically structured human organizations. Those in authority direct, command, supervise, guide, and manage, those who are subject to them. As we have seen, Plato develops this analogy in the *Republic*, drawing a parallel between top-down imposition of order in the mind and the state. The self-intervention techniques described in the last section involve the agent in a quasi-managerial or supervisory role.

According to the picture of the mind emerging from contemporary psychology, conscious control sits atop a vast bureaucracy of semi-independent mental agencies (Wilson 2004). What the activity of self-governance opens up is the possibility of deploying processes of control to manage

this vast, semi-autonomous infrastructure in light of goals deemed normatively worthy. Our desires, affects, and attitudes tend to have a life of their own. This is not to say they are always unruly and opposed to normative judgment, only that there is for us often a gap between judgment and our affective and motivational life. Bringing them into harmony requires some work; we have to *exert* some effort and *impose* control. Because the agencies of our mind are semi-independent, being only partly responsive to judgments issued by the conscious deliberating and choosing part, the control we wield over our minds is frequently indirect.

The activities described in the last section include things like planning, control of appetites, regulation of affect and attitude, and habit management. By engaging in this kind of self-intervening activity, agents can sway affective, motivational, and other attitudinal states that are not subject to direct control. In short, self-management recruits elemental control abilities to spread control to a wider sphere of life.

4. Consciousness

The phenomenon of self-governance I have highlighted involves high-level cognitive control characterized by processes that are conscious, and frequently deliberative and effortful as well. What warrants this focus on higher level conscious processes?

There has been an expansion in recent philosophy of agency, from the traditional focus on conscious deliberated action to a wider view that emphasizes the importance of unconscious, habituated, and skilled agency (Doris 2015, Montero 2016, Railton 2009). This expansion is salutary. It enriches the picture of human agency and helps make it more realistic. My emphasis on high-level conscious processes is not meant to deny this broader picture. High-level conscious

processes are continuous with, and dependent on, lower-level and unconscious self-regulatory processes. As a rich literature in the psychology of agency shows, goal-directed action depends on various controlled-processes that are frequently unconscious. If we think of controlled processes along a spectrum of more or less conscious (Braunstein et al. 2017, Churchland & Suhler 2009, Cohen 2017), with fully deliberated and self-aware processes at one end of the spectrum, then the processes I have described are toward that end of the spectrum. Unconscious influences on action have been widely documented (Bargh 2017, Wilson 2004). Assuming conscious processes also make a difference to action (Mele 2014), then there appears to be causal influence going both ways, which we might picture as an information highway connecting the two ends of the spectrum. Conscious processes are typically "higher up" in the control chain of action, not because they are immune to lower-level influence, but because they tend to structure and guide lower-level processes.

The conscious deliberative perspective holds special importance for our self-understanding as agents and is key to understanding the phenomenon of self-governance I am after in this chapter. My focus on conscious processes is motivated by an interest in understanding the special form of agency involved in personal autonomy. It is not meant to suggest an unrealistic picture of the mind, skewed toward the conscious and the deliberative. On the contrary, it is meant to suggest a picture of the mind on which it is *not* skewed toward the conscious and the deliberative, but rather one on which there is room for both kinds of processes to work together. The self-governing agent I have described is an embodied rational agent, only partly aware of the motives that drive her, and only in a fragile relation of control to large parts of her life. My interest is precisely in how finite, embodied agents like us can manage our fragilities and work within our limits. This requires some

awareness of those limits and creative strategies of self-stewardship which can be consciously controlled and implemented.

5. Self-Governance under Norms

Executive self-regulation is not a uniquely human trait. For example, in order to gain a larger reward, pigeons (not paragons of self-control in the animal kingdom) can use a learned pre-commitment device wherein they peck at one button to keep themselves from pecking at a second button (Ainslie 1974). Closer to phylogenetic home, chimpanzees have been shown to succeed at a variety of delay-of-gratification tasks. In an interesting parallel to Mischel's findings for young humans, chimpanzees play with toys to strategically distract themselves from going for the tempting proximate option in a delay-of-gratification task (Beran 2015). However, the possibilities for executive function do ramp up considerably with increased brain-size. MacLean et al. (2014) tested response-inhibition on the same two tasks across 36 species, including birds, elephants, Canidae, lemurs, and old- and new-world monkeys, and found that absolute brain size was the single best predictor of self-control across species.

The large human neocortex enhances cognitive control abilities and puts at our disposal a more sophisticated repertoire of cognitive tools for intentional self-management. But novel psychological structures in humans also fundamentally transform the possibilities for self-regulation. Human norm psychology makes possible normatively articulated self-governance: once the evaluation of oneself and others in light of norms comes on board, capacities for self-governance can be deployed in the service of conformity to norms. This in turn is partly mediated by distinctive human metacognitive abilities: we can think about and reflectively evaluate our own

mental states—our beliefs, intentions, desires, and emotions. Many philosophers have seen these twin capacities for metacognition and evaluative thought as central to distinctively human agency (cf. Brink 2008, 2019; Frankfurt 1971, Korsgaard 1996, 2009; Watson 1975). Moreover, adding to this suite of abilities, humans can share perspectives (Tomasello 2019), engage in justificatory exchange (Pettit 2001, Sperber & Mercier 2017), and view themselves both as subjects and objects (Nagel 1989). This makes it possible for humans to entertain justification-questions about the appropriateness of their mental states and, more radically, about the adequacy of the norms in light of which they reflectively evaluative and guide their behavior. This turns humans from mere self-controllers into self-governing normative agents.

Human capacities for self-regulation are thus continuous with, and depend on, self-regulatory capacities found throughout the animal kingdom. But while they build on our mammalian and primate inheritance, specifically human abilities for self-governance introduce novel features. In particular, thanks to self-consciousness and norm-psychology, humans are able to deploy their self-governance capacities in the service of normatively articulated ends.

6. Self-Governance and Responsiveness to Reasons

What is the relationship between the activity of self-governance, as I have described it, and reasons-responsiveness? Reasons-responsiveness consists in recognition *and* conformity to reasons. We can therefore distinguish between what Fischer and Ravizza (1998) call the receptivity and reactivity aspects of reasons-responsiveness, or what Brink and Nelkin (2013) call its cognitive and volitional dimensions. Let's consider each of these in turn.

Cognitive. Automatic, skilled, and habituated behavior can surely be reasons-responsive (cf. Railton 2006, 2009). Perhaps much of the time we recognize reasons in a way analogous to fluid and automatic perceptual processing. Most people can navigate a busy street, negotiating traffic, avoiding pedestrians, stepping aside for the elderly person with the walker, and so on, all while being absorbed in thought about other things. Likewise, most people can recognize without any reflection that they have reasons not to step into the street, to avoid the oncoming pedestrian, and to pause for the elderly individual. But while such automatic and effortless "perception" of reasons is indispensable for human normative agency, it clearly isn't the whole story. We also often find that we don't know what we have most reason to do. In many circumstances, we find ourselves not able to simply perceive reasons the way we perceive objects in our environment; instead, we find ourselves needing to figure out what reasons we have or how to weigh competing reasons.

On the cognitive side, then, we often have to pause and deliberate to appreciate what reasons we have or what they indicate. What bears emphasizing here is the way in which deliberation can be an *indirect* self-governance strategy of the sort described earlier. In extended reflection, one does not automatically appreciate reasons but has to take steps to put oneself in a position to do so. One attempts to bring to mind relevant considerations. This requires retrieving items from memory and assessing their relevance. (Has one thought of all the relevant considerations, or at any rate the most significant ones? Is this consideration really relevant?) One then needs to ferret out the strengths of these considerations in relation to other considerations to arrive at an overall verdict. (Which of these two competing considerations is more important in the present context? If one is more important, does adding a smaller contributing consideration on the other side tip the scales?) One might also need to engage in activity designed to promote the

quality of one's deliberative process. One may need to take a step away from a situation and emotionally re-calibrate, find a reflection-conducive space or atmosphere, remind oneself of various things (e.g., that one is prone to self-deception about certain matters), use one's imagination in a variety of ways (e.g., to project oneself into another person's shoes), find a conversation partner who can offer helpful input or simply act as a reflective "sounding board," and so on.

This entire process can be vexed and arduous, especially when the stakes are high. And it is precisely necessary at times when we *do not* directly and immediately appreciate our reasons. Instead, we have to engage in activity in order to help us discern them. To be sure, the attempt to discern reasons will not get very far without some antecedent attunement to appropriate norms and values. The enterprise of seeking reasons would surely be a lost cause unless there is sufficient antecedent sensitivity to reasons. But my claim is not that self-governance activity is sufficient for reasons-responsiveness, nor even that it is necessary for all reasons-responsiveness. As I said, we operate on automatic pilot much of the time in responding to reasons. My claim, rather, is that self-governance is necessary for *some* reasons-responsiveness. In particular, it is necessary for the full-orbed reasons-responsiveness that characterizes normal adult functioning. Under realistic conditions, to achieve high or even moderate reasons-responsiveness in the course of an ordinary human life requires some work on the part of the agent to make herself "receptive" to reasons. Above I described how complex self-interventions utilize more basic controlled activity. Likewise, the process of attempting to deliberatively discern reasons presupposes more basic attunement to normative features of the world. Such basic attunement will not be magically produced by acts of deliberation, but acts of deliberation can facilitate, enhance, and deploy those basic capacities.

Volitional. Imagine humans who experienced no discrepancy between their behavior and their normative commitments. Always and without fail, these humans do whatever it is they judge good and right. Suppose further that this remarkable coincidence results from the fact that they experience no motivational or habitual opposition to their normative judgments, and that they inhabit environments that never bring them off course. They effortlessly behave in the ways they think they ought to. Ordinary humans are not like this. Our normative evaluations frequently diverge from our behavior and we often find it to some extent effortful to make our behavior conform with our evaluations. We feel the tug and pull of contrary motivation; the lag of habituation; the sabotaging effects of circumstance. Under normal conditions, then, we have our work cut out for us: we must find ways to creatively and constructively manage ourselves to bring our lives into line with our evaluative commitments.

Self-governance activity can be deployed in sundry ways to help us conform to the reasons we recognize. When there is a gap between our normative judgments and our attitudes, motivations, and actions, self-governance activity can be a way of producing greater alignment. Peter judges that he ought to take more time off work, play and relax more, and spend more time with his wife and children. But he finds his lifestyle oddly out of sync with his judgments. Recognizing this, Peter finds ways to strategically implement changes: he sets a daily visible reminder to stop working at a particular time, joins a soccer league to "nudge" himself to do more fun recreational activities, schedules a regular date night with his wife, and pre-commits time to his children through advance scheduling. As we saw earlier, self-governance interventions may also need to target attitudes and emotions.

In sum, the activity of self-governance is plausibly a way we exercise, realize, and achieve (fuller) reasons-responsiveness. It can play a role, both in helping agents come to appreciate the reasons they have and in conforming their behavior to those reasons.

7. Self-Governance and Structural Virtue

To pursue a bit further the connection between self-governing activity and the kinds of substantive goods I have placed at the center of autonomy, it is useful to think of self-governing activity as closely analogous to (and perhaps a species of) what Robert Adams calls a "structural virtue." Adams (2006: 33-34) writes:

> I say that capital V Virtue is persisting excellence in being for the good. One implication of this is that in ascribing Virtue, holistically, to a person I must in a general way commend her being for what she is for and against what she is against. But not all the particular virtues are essentially ways of being for and against things one should be commended for being for and against.
>
> Some of them are. Some virtues are defined by motives which in turn are defined by goods that one is for in having them, as benevolence, for example, is defined by the motive of desiring or willing the good of others. We may call these motivational virtues. They would not be virtues if the ends they are definitively for were not goods, and goods that it is in general excellent to be for.
>
> Other virtues—courage, for example, and also self-control and patience—are not defined in that way, by particular motives or by one's main aims, but are rather structural features of the way one organizes and manages whatever motives one has. We may call these structural virtues. The excellence of structural virtues is a matter of personal psychic strength—of ability and willingness to govern one's behavior in accordance with values, commitments, and ends one is for. However excellent they may be as strengths, structural virtues by themselves cannot make one a morally good person. That depends above all on ''having one's heart in the right place,'' on what goods one is for, and thus on motivational virtues. But without some of the strengths of structural virtues one can hardly be excellently for the good.

As I interpret Adams, the key distinction between capital V Virtue and structural virtue is that the first is substantively defined whereas the second is not. Capital V Virtue is defined in terms of

what is actually good. If someone is deeply misguided in her "benevolence," then she doesn't manifest the virtue of benevolence but its misguided analog. To count as genuine benevolence, the agent's attitudes and actions must track and be appropriately responsive to the (genuine) good of others. Structural virtue, by contrast, is not defined in terms of what is actually good. One might be a courageous warrior fighting for a wicked cause, a disciplined thief, a self-controlled miser, a circumspect sadist, a person who is patient with injustice, and so on. Nevertheless, because of our characteristic limits and frailties, structural virtues are general agency assets in human life that play a vital role in enabling and sustaining human excellence. Structural virtues are not inherently good (they can be neutral when lives are organized around worthless ends or bad when lives are organized around evil ends), but when they are appropriately aligned with, and in the service of, substantive virtue, they can be very good. In this complimenting or supporting role, they are in fact indispensable. To be effective and enduring traits, virtues like beneficence and justice need structural virtues to support appropriate agential functioning against a variety of inner and outer obstacles.

The activity model of self-governance, as I have been exploring it, bears close resemblance to Adams' notion of structural virtue. Self-governance, as I have unpacked it, consists in active self-management, in the agent's doing things that allow her to deploy elementary capacities to expand the sphere of control, including to deal with quintessentially "passive" states like affect and desire. Such self-management strategies are not defined by, and can come apart from, reasons-responsiveness. Instead, they are defined in value-neutral terms by the functional role they play (extending control) and as being in the service of, and effective relative to, some specification of goals—*whatever those goals are*. Hence, self-governance in this sense is purely formal. As such, it is a lot like structural virtues, which can support neutral or evil activity.

I suggested above that self-governance does not guarantee reasons-responsiveness. Without a modicum of antecedent sensitivity to relevant normative features, cognitive and volitional self-stewardship will be powerless to make one responsive to reasons. We can add a further point here, viz. that there is no guarantee that self-governance will secure or amplify reasons-responsiveness, even in those who have the basic antecedent sensitivities. After all, even if one aspires to discern and respond to the reasons one has, one can make mistakes and end up deeply misguided. As I see it, then, self-governance in the sense expressed by the activity model is compatible with being quite deeply off-track, normatively speaking. A thoroughly evil person might be self-governing, as might a person who, though not evil, is drastically out of touch with (moral or prudential) reasons. Thus, the person dedicated to counting blades of grass (Rawls 1971/1999: 379) or collecting lint (Brink 2008: 24) might be a virtuoso self-governor in this sense, though quite out of touch with reasons.

Nevertheless, like structural virtue, active self-management is indispensable in the ordinary course of life and plays an important role in genuinely worthwhile human achievement. While self-governing as a purely procedural phenomenon neither presupposes nor guarantees reasons-responsiveness, it does play an important role, both in the exercise and enhancement of reasons-responsiveness for finite agents like us. Unlike the imaginary humans referred to above, real humans have to work at being reasons-responsive. That is of course not to say that we don't sometimes, even often, respond to reasons on autopilot. I have already granted that we do this. Rather, it is to suggest that in the ordinary course of life, to be adequately responsive to reasons, and certainly to be optimally responsive to reasons, requires agents to engage in self-governing activity.

Engaging in self-governing activity is, at least in part, how we exercise our responsiveness to reasons. Thus, if we assume Valeria has general capacities for reasons-responsiveness, and that she possesses these capacities in respect of (most) situations in which she has opportunities for playing the violin, part of what it means for her to have the capacity is that she can perform suitable cognitive and volitional self-interventions to get herself to play the violin; and on many occasions (e.g., when there is some motivational recalcitrance) part of what it will mean for her to successfully realize those capacities is engaging the relevant cognitive and volitional operations to move herself to appropriate action. Engaging in self-governing activity is also, at least in part, how we enhance our responsiveness to reasons. Perhaps Valeria enjoys moderate responsiveness to reasons. Self-governing activity can help improve this. Suppose that, without engaging in much self-governing activity, Valeria is able to respond to the reasons she has to play violin with a modicum of success. By actively engaging herself, however, Valeria is able to bring it about that she responds to those reasons even better.

The parallel between self-governing activities and structural virtues, then, should be relatively clear. Structural virtues can help to bring about and sustain capital V virtues, but they don't entail those virtues. Because capital V virtues are modally robust and enduring, however, capital V virtues do entail having the requisite structural virtues to sustain them. Similarly, active self-governance can help to bring about and enhance reasons-responsiveness, but it doesn't guarantee reasons-responsiveness. Conversely, however, under realistic conditions reasons-responsiveness does entail requisite formal abilities for managing one's life. The main difference, it seems to me, is that structural virtues are traits of character, whereas self-governance activities, though presumably supported by traits of character, are not themselves traits of character but rather dynamic processes.

Conclusion

This chapter began by comparing three models of self-governance and then explored one of these models, the activity model, according to which people govern themselves when they engage in activity that has as its aim the regulation of behavior. I explored a variety of aspects of self-governance in this sense and concluded by comparing it to the notion of a structural virtue.

In spite of my focus on the activity model of self-governance in this chapter, the reason-first view of autonomy developed in the last chapter can be comfortably ecumenical about the language of self-governance. One might interpret self-governance, as the rule-of-Reason model does, in terms of successfully responding to reasons, in which case self-governance would align with what I have called perfection-entailing freedom. Alternatively, one might interpret self-governance in a sense tracking the possession of normative capacities rather than their realization, in which case self-governance would align with that I have called responsibility-entailing freedom. Either way, self-governance talk would correspond to concepts I have argued are deeply implicated in our thinking about autonomy and cannot be lightly dispensed with. Moreover, while the reason-first view does privilege the idea of rational control, it by no means excludes concerns about authenticity. Consequently, the reason-first approach should have no complaint about using self-governance talk in a way meant to track whether agents are operating from authentic attitudes and preferences. The language of self-governance, as I say, is flexible.

My reason for focusing on the structural notion is that it is vital for understanding how substantive autonomy is realized. Under ordinary conditions, rationally limited and fragile agents like us must work at being autonomous. Naturally, the process of self-governance as described in

this chapter is no magic bullet. It doesn't guarantee reasons-responsiveness or secure us against the possibility of a hijacked self. Nevertheless, under ordinary conditions, it is part of how we exercise and achieve our autonomy.

Chapter 5

The Value(s) of Personal Autonomy

Over the course of the preceding four chapters, I have been building a case for the attractiveness of a certain way of thinking about autonomy, one which recognizes a plurality of distinct elements that can be fruitfully held together and illuminated by putting the idea of rational control or reasons-responsiveness at the center. Along the way, I have appealed to claims about autonomy's value. In particular, I have argued that autonomy's value must be fit to play a dual normative role, grounding both reasons to respect it and reasons to promote it. An account of autonomy's value should be consistent with this normative profile, illuminating why autonomy is a value to be respected and promoted. Formal views, I have argued, have difficulty vindicating this twofold normative role. By excluding reasons-responsiveness as an essential ingredient, they leave it dubious whether autonomous agents are responsible for their choices in a sense sufficiently robust to ground strong presumptions against interference with their choices, while at the same time making it more challenging to see why autonomy is a desirable agency ideal worthy of promotion. The reason-first view, by contrast, is better equipped to vindicate autonomy's dual normative role, adding a crucial pillar to the basis of respect and providing a richer set of resources for recovering autonomy's value as a worthwhile agency ideal.

This chapter looks at autonomy's value in more detail. One of my central claims has been that our thinking about autonomy is complex and that we do well to honor this complexity while seeking a theoretical framework that allows us to simultaneously preserve and unify different conceptual elements. In sections 1 through 5, I therefore proceed by asking about the value of each

of the different conceptual strands introduced in earlier chapters. Section 6 concludes by stepping back and attempting to fit these considerations into a larger frame.

1. Perfection-Entailing Freedom

It is unclear to what extent perfection-entailing freedom is or is not part of the common-sense repertoire, that is, of the way people naïvely or pretheoretically conceive of autonomy.[45] Nevertheless, I argued that perfection-entailing freedom turns out (perhaps surprisingly) to be a commitment internal to our thinking about autonomy, since it helps account for autonomy's full value as an agency ideal worthy of pursuit and promotion (chapter 2), and since it coheres deeply with our own rational ends as normative agents (chapter 3).

There are at least two plausible candidate answers to the question, What is valuable about perfection-entailing freedom? The first is that perfection-entailing freedom is valuable because it consists in responsiveness to substantive (practical) reasons—nothing more.[46] Say that Ayumi has reasons to lead the company she has started, to be a good parent to her two children, to spend time with her friends, to learn about the history of Japanese art, and so on. What is valuable about Ayumi's succeeding in responding to her reasons? Just this: in doing so, she realizes the values contained in these activities—leading a business, raising her children, spending time with friends, studying art history, and so on.

Of course, for this answer to be plausible, (practical) reasons must be closely tied to substantive value (Raz 2003, Scanlon 1998). This is something I have assumed in earlier chapters.

[45] Thanks to Dana Nelkin for pressing me on this.
[46] My focus here is on responsiveness to *practical* reasons. As I suggested in chapter 3, responsiveness to epistemic reasons will also be part of full autonomy. I leave aside how to think about the value of responsiveness to epistemic reasons.

As long as this picture of the connection between reasons and values is correct, the "nothing more" answer seems elegant in its simplicity. Perfection-entailing freedom is valuable because of all of the particular values that are realized in perfection-entailing freedom. The value of perfection-entailing freedom reduces entirely to those other values. Nothing more is needed.

A second answer is to say that in responding successfully to one's reasons, one thereby realizes some further good. There might be different ways to specify this additional value. The suggestion I want to consider is that responding to one's reasons manifests a distinctive excellence, which is valuable in its own right. T.H. Green (1886/1996) gives powerful expression of this type of view. On Green's way of thinking about it, autonomy is identified with self-determination by reason, where this is contrasted with mere responsibility-entailing freedom (233) and requires the actualization of rational potential (244). Green interprets this actualization in the language perfection (245) and self-realization (246). Since reason is one's own highest self (234), the rule of reason is no alien principle but a realization of one's own nature and truest self.

Perfectionism, as it is usually understood, is a doctrine about welfare or the personal good. It says that welfare or the personal good consists in realizing human nature. On some views, human nature is understood as a biological kind (Hurka 1993). On others, it is understood as a normative kind, that is, human nature is interpreted in terms of a conception of persons as rational agents (Brink 2008, 2019). At its most ambitious, perfectionism claims that the rich tapestry of human values can be organized around, and explained in terms of, the perfection of our (rational) nature. For our purposes, what matters is a weaker claim, viz. that the perfection of our (rational) nature is *one* important good among others.

Within the reason-first approach, persons are conceived as normative agents with capacities for substantive practical reason. I argued in chapter 3 that persons as rational valuers are

implicitly committed to the aim of practical reason. Given their capacities for appreciating and responding to genuine values and reasons, and their rational aim of successfully exercising their capacities, perfection-entailing freedom would (as the name is meant to suggest) plausibly count as a valuable kind of realization of their essential nature as persons—that is, a perfection of their normative agency (Brink 2019). In successfully exercising their normative capacities, persons therefore realize a distinctive kind of human excellence or virtue.

This excellence is not reducible to the many particular substantive values realized in perfection-entailing freedom. The claim is not that this excellence is disconnected from the many particular values that are realized. On the contrary, it is deeply bound up with the successful pursuit of particular values. Rather, the claim is that this excellence has value in its own right and doesn't depend on any particular pattern of instantiation. It is the same for other virtues. Think of the virtue of benevolence. Benevolence is deeply bound up with the successful tracking and pursuit of one kind of value—roughly, welfare. But while an agent cannot possess the virtue without exercising benevolence on particular occasions and toward particular individuals, the virtue is not reducible to those exercises, since it does not depend on any particular pattern of instantiation.

To illustrate: Bongani has the virtue of benevolence. He cares robustly about others—especially the poor, the abandoned, and those in need—and he marshals his resources to help when he can. He adopts an abandoned young boy and tutors him, cares for his elderly mother, and sees to it that the poorest villagers never go without food. Bongani's benevolence is, of course, deeply bound up with the valuable goods and activities that fill his life. But while his particular excellence as an agent is *displayed* in his life's particular engagements, it is not reducible to them. Bongani might not have met and helped this particular boy. Instead, he might have met and helped a

different boy. His mother might have died much earlier and not required the care that she did, in which case, perhaps Bongani would have adopted a second child and cared for it.

Something similar goes for the excellence that consists in responding to one's reasons. Insofar as persons realize a distinctive kind of human excellence in responding to the reasons they have, the value of that excellence could be equally realized by responding to different reasons. Though two lives differ in the particular pattern of activities and choices, they can manifest the same excellence. Ayumi's excellence in successfully pursuing the reasons she has is, of course, deeply bound up with the particular values that fill her life—entrepreneurship and leadership in business, the lives of her children, the value of Japanese art, etc. But Ayumi might have partnered with a different man and therefore had different children or decided against partnership and childrearing, might have become an historian of Japanese art rather than a business entrepreneur, and so on.

In philosophy jargon, the excellence displayed by Bongani and Ayumi is multiply realizable. Their excellences depend on some appropriate pattern of instantiation, but not on any pattern in particular. Of course, the excellence that consists in successfully responding to one's reasons is realizable in a far wider set of patterns than benevolence (or any other special virtue). Benevolence tracks a particular kind of value. Responsiveness to reasons does not track any one kind of value. Assuming a plurality of values, there are thus many more ways to be responsive to reasons than to be benevolent.

It is plausible, then, that responding to one's reasons manifests a kind of agential excellence whose value is not the same as any particular pattern of responsiveness to reasons. This agential excellence is, of course, a highly abstract and generic kind of excellence. Indeed, we might say it is the most abstract and generic kind of excellence of which humans are capable. This may make

it seem like a rather boring and bloodless virtue. But that impression derives from its abstractness rather than its value. Since there are many different values and many ways of appropriately responding to those values, the excellence of being successfully reasons-responsive must, of necessity, be a rather high-level and abstract virtue. It should be clear that this is no knock against the value of being successfully responsive to reasons; it is to ascend upward in abstraction to achieve the widest characterization of virtue possible. Succeeding in responding to one's reasons is just the widest specification of excellence for normative agents.

Now Green seems to me right to think that this type of generic rational excellence is also a kind of *self*-realization or *self*-actualization, not merely a kind of species-perfection which happens to be located in individuals. Notice three things. First, assuming persons are essentially normative agents, then being responsive to reasons is also a realization of *their* nature *as individual persons*. Because Ayumi is fundamentally a normative agent, succeeding at responding to her reasons as she does is a kind of self-realization—a realization of what *she* fundamentally is.

Second, if agents are committed to the aim of successfully deploying their normative capacities, then their success in doing so is not an alien imposition, a value imposed on them from without with which they might not be identified. The aim of succeeding in responding to her reasons is also one which Ayumi is herself committed to. So she cannot feel alienated from this value as merely some abstract good of species-perfection: it is tied to her own aims as a rational agent.

Third, if normative space is frequently permissive and agents create valuable lives by selecting among permissible options, then there is considerable room for individual uniqueness in the development of the generic value. The particular way in which Ayumi manifests her rational excellence bears her own special imprint. Over the course of her life, she makes many choices that

give her life a unique profile and reflect her own creative input. Had she decided to become an historian of Japanese art and not to have children, her life would look very different than it in fact does. The reason her life takes the particular shape it does is that *she* made the particular choices she did. This would be true for any particular instantiation of the generic value, i.e., even if Ayumi had made different choices (consistent with responding to her reasons). Hence, there is a further sense in which successfully responding to one's reasons can be a kind of *self*-realization: the particular pattern of one's choices can bear the agent's distinctive imprint.

2. Responsibility-Entailing Freedom

Since responsibility-entailing freedom is presupposed in perfection-entailing freedom, it is at least valuable *as a condition* of the latter. The value of responsibility-entailing freedom more generally seems to me like this. It has value because it is a condition or amplifier of other things of value.

There are a range of human goods which in some way depend for their value or significance on background conditions of free and responsible agency. For example, achievements typically depend on their being freely and responsibly undertaken. If one is coerced into climbing a mountain, the successful completion of the task may exhibit skill and athleticism, but it loses value as an accomplishment.

The value of relationships can be similarly sensitive to background conditions of free and responsible agency. The question of whether, and if so, what kind of freedom adds value to symmetric relations of friendship and romance is complicated (Kane 1998, Pereboom 2014). A relatively modest claim suffices for the present point: background conditions of free and

responsible agency can *sometimes* contribute to the *full* value of relationships. I take it that the ideal of intimate love and deep friendship speaks against factual and normative delusion. Ideally, we want our lovers' affections to be based on a reasonable appreciation of what we are actually like, including our *genuine* merits. So while it is true that love is often experienced as passive, there is still a significant difference between love and compulsion (cf. Frankfurt 1999). My claim is that background conditions of free and responsible agency can contribute to valuable forms of love.

The good of meaning in life is plausibly also conditioned by free and responsible agency (G. Dworkin 1988: 20; Wall 1998: 147). I suggested in chapter 2 that at least part of the appeal of metaphors of self-creation and self-authorship is that they express in rather vivid imagery the thought that autonomy is responsibility-entailing. According to Raz (1986: 369, 390): "The ruling idea behind the ideal of personal autonomy is that people should make their own lives. The autonomous person is a (part) author of his own life. The ideal of personal autonomy is the vision of people controlling, to some degree, their own destiny, fashioning it through successive decisions throughout their lives…Personal autonomy is the ideal of free and conscious self-creation." The exercise of self-creation or self-authorship is meaningful for creatures like us. This is presumably because one of the deepest truths about us is that we are agents, not just passive experiencers of the world. It enriches and gives meaning to life when we engage it actively, freely, and responsibly.

Part of why responsibility-entailing freedom is valuable, then, is that it either enables or enhances a range of important human goods. It functions as a gateway good or as an amplifier of value. But what if these further goods remain unrealized? Does normative capacity have value as such, apart from its role in securing or enhancing other goods? It is worth refining this question by distinguishing two sorts of value, one involved in welfare and the other in recognition respect.

The first sort of value contributes (constitutively or instrumentally) to an agent's welfare. The second sort of value grounds reasons to treat an agent with recognition respect (cf. Darwall 2006). Since responsibility-entailing freedom consists in enjoyment of normative capacities sufficient to make one robustly responsible for one's choices, it seems clear enough that it has value of the second sort. The crucial question concerns the first sort of value. Does the mere possession of normative capacities (plus opportunities) enhance an agent's welfare?

It is not clear that it does. Recall Randy the gamer. Randy squanders opportunity and talent playing endless hours of video games. Let's stipulate that, given its role in Randy's life, playing video games is not welfare enhancing. (Video game playing might have value as an occasional leisure activity or temporary diversion, but since it is an intrinsically worthless or nearly worthless activity, it can merit neither a great deal of investment nor serve as proper organizing value for a life.) Does the fact that Randy opts for video game playing in full possession of normative capacities somehow add to his welfare? As noted, we can appreciate the significance of Randy's normative capacities for what it means to respect him. But this is different from recognizing their value as welfare-enhancing. By hypothesis, playing video games does not contribute to Randy's welfare. It is therefore difficult to see how the mere fact that his choices issue from responsibility-relevant capacities makes his life go better for him.

The point is general. Since capacities sufficient to ensure responsibility are compatible with making bad choices, including choices an agent herself would regard as detracting from her welfare, the mere possession of normative capacities does not as such appear to contribute positively to welfare. Indeed, when a choice is very bad, an agent's making it freely and responsibly might even make her life go worse for her.

Yet because responsibility-entailing freedom makes perfection-entailing freedom possible, and because more generally it makes some significant human goods possible and enhances others, it is nevertheless *prospectively* valuable, whatever agents ultimately make of their freedom. Fred Feldman's (2004) "crib test" is a convenient way to get a handle on the kind of value that is at stake in welfare. We are to imagine parents who want the best for their child. What they want is that the child's life goes as well as it possibly can *for her*. Now if what I have said is plausible, parents should want their children to enjoy responsibility-entailing freedom because of its value as a gateway good and value-enhancer. Having such freedom will make it possible for the child to live the richest and most valuable kind of life it can. Suppose the parents have the choice to either endow their child with such freedom or withhold such freedom. At the moment of choice, they do not know how their child will use its freedom. They therefore do not know the precise contribution such freedom will make to enhancing their child's welfare. Nevertheless, it would be rational for the parents to endow their child with responsibility-entailing freedom. While its ultimate contribution to the child's welfare depends largely on what the child goes on to make of her freedom, considered prospectively or *ex ante* it would be rational to choose such freedom for her.

In sum, whereas perfection-entailing autonomy guarantees positive value, responsibility-entailing autonomy does not. Considered in terms of its contribution to welfare, responsibility-entailing freedom is largely valuable because of the goods it makes possible.

3. Authenticity

A distinction was made in chapter 3 between two different roles played by the idea of authenticity within a reason-first approach to thinking about autonomy. The first is that recognition

respect is owed to persons as particular individuals, and an account of authenticity can help clarify what this concretely means. The most general ground of recognition respect for persons is generic: it is a person's nature, the fact that she is a being of a certain kind, that calls for treating her in a way that befits her status as a being of that kind. But in many situations, adequately respecting a person requires taking into account more specific facts about her—her proclivities, sensibilities, desires, hopes, values, beliefs, commitments, and so on. Some of these will be more central than others to a person's evaluative outlook and sense of identity. By mapping depth and structure within the self, an account of authenticity can thereby help clarify at least part of what is involved in respecting persons as the particular individuals they are, so that—to borrow Nandi Theunissen's (2018: 367) apt way of putting it—" we relate to them always with a view to their being the center of a life to which they bear a special relation." In short, there are both general grounds of respect and particular determinants of respect. The former pick out generic features in virtue of which recognition respect is owed; the latter pick out particular features which determine the distinctive shape respectful treatment must take. An account of authenticity can help clarify the particular determinants of respect.

The second role played by the idea of authenticity is as an agency ideal. In this role, authenticity is something we have reasons to seek for ourselves and promote the realization of in others. A full exploration of authenticity as an agency ideal is beyond the scope of this chapter. However, it is worth making three points about authenticity as an agency ideal within a reason-first view of autonomy.

First, there plausibly are a variety of agency ideals in the vicinity, ideals like wholeheartedness (cf. Frankfurt 1988), integrity, and individuality. Considerations about value will therefore need to begin by getting clear about the particular authenticity ideal in question.

Second, some of these ideals may have relatively basic and non-instrumental value. Take individuality. In *On Liberty* John Stuart Mill (1859/2003: 131) writes: "There is no reason that all human existences should be constructed on some one, or some small number of patterns. If a person possesses any tolerable amount of common sense and experience, his own mode of laying out his existence is the best, not because it is the best in itself, but because it is his own mode. Human beings are not like sheep; and even sheep are not undistinguishably alike." Why is individuality valuable? Presumably, individuality is valuable for persons because something about their nature makes it so. Humans are not sheep. But one reaches explanatory bedrock pretty quickly here, at any rate, if one is focused on explaining individuality's non-instrumental value.

We can nevertheless unhesitatingly affirm the non-instrumental value of individuality. Though then it looks like there can be tradeoffs between the values of individuality and responsiveness to reasons. Individuality and responsiveness to reasons need not, of course, conflict most of the time. But what about when they do? What about cases in which more individuality means less responsiveness to reasons? One option is to say that the value of individuality is strictly conditional on harmonizing with responsiveness to reasons so that there can never be genuine conflict between the two: when individuality is opposed to responsiveness to reasons it has no genuine value. Another (and it seems to me more plausible) option is to say that the value of individuality is not strictly conditional in this way and that there can, therefore, be tradeoffs between individuality and responsiveness to reasons.

Third, while the reason-first approach need not dictate any particular interpretation of authenticity ideals, it does provide a framework within which they can be interpreted and their value assessed. Even if authenticity ideals, like individuality, have *some* independent value, they need not have much independent value considered apart from responsiveness to reasons.

This has implications for trade-offs between authenticity ideals and responsiveness to reasons when they clash. All else equal, it is desirable to be wholehearted rather than deeply ambivalent, but if psychic harmony insulates one from recognizing or responding to reasons one has, it may be better to be perturbed and divided. All else equal, it is desirable to have and act in conformity with one's principles rather than not to have any principles or not to act in accordance with them, but if one's principles are substantively misguided, it may be better to act against them (cf. Arpaly 2002). All else equal, it is desirable to cultivate one's individuality, but if one's project of forming a distinctive self becomes unmoored from what is actually good for one, it may be better to be less distinctive.

Within a reason-first view, responsiveness to reasons is the organizing and guiding value. One need not deny that authenticity ideals have some independent value which might at times conflict with responsiveness to reasons. Instead, a better way to think about the relationship between authenticity ideals and the reason-first framework is to say that these agency ideals are regulated by a deeper agency ideal, one which is given explanatory and evaluative priority. Hence, when authenticity ideals clash with responsiveness to reasons, they will usually not win out, since they have at best minimal independent value. Moreover, within the reason-first view, the fact that agents experience misgivings, doubts, irresolution, psychic division, and even betray their principles, can be a very good thing. Finally, it is only when authenticity ideals like wholeheartedness, integrity, individuality, and so on, become integrated with our substantively rational aims that their value as aspirational goods—as things that are genuinely desirable—becomes fully secure.

4. Independence

Chapter 1 distinguished two kinds of independence, one external and social, the other internal and attitudinal. Let's briefly consider the value of each.

The case for the value of external independence is fairly straightforward. External independence is valuable because of the various values it protects and promotes. In an autonomy-favorable milieu, people can develop and exercise their normative capacities; they can pursue valuable forms of authenticity, engaging in valuable projects of self-formation to become distinctive selves; and they can engage in valuable projects of self-creation, crafting lives oriented around a variety of genuine values. External independence secures the possibility for these goods. By securing strong protections for individual choice, it will also tend to encourage people to exercise and develop their rational agency capacities.

External independence may also be constitutively valuable. Arguably, perfection-entailing freedom depends constitutively on a suitable external environment. If we side with Aristotle and Mill against Bentham in thinking that the human good contains a large active ingredient, then we will think it matters not only that people arrive (however they do so) at their ends but that they play the role of an agent—an active, engaged, creative, deliberative role—in getting there. There could plausibly be genuine goods in an engineered world in which people need not much engage their normative capacities for deliberation and choice, and in which their powers for creative willing are largely dormant. One can imagine people being funneled into lives that are (in some sense) good for them. But in such a world, there would be an unimaginable loss of value and, more to the present point, that value seems *constitutively* ruled out in such a world, ruled out by the external controlling and canalizing setup. For it to be the case that one's life and choices are the meaningful upshots of exercises of one's own agency, external freedom is required.

The value of internal independence is less straightforward. Much depends on matters of detail. Which type of attitudes are we talking about? Just what do we mean by independence? In which context? And so on. I won't investigate these complexities here. What I want to briefly focus on instead is the general value of being resistant to conformity and resilient against pressure to change one's mind.

An important body of literature in social psychology, beginning with the pioneering work of Solomon Ash (1955) and Muzafer Sherif (1965) documents the profound susceptibility of humans to conformity effects. And the celebrated obedience studies of Stanley Milgram (1974), which have recently been replicated (Burger 2009), show how susceptible people are even in liberal societies to heeding perceived authorities.

These deferential tendencies can lead to socially and personally sub-optimal outcomes. As Mill argues in *On Liberty* (1859), the tendency to conform can stifle dissent and the discovery of new truths. More recently, Cass Sunstein (2019) has argued that it can give rise to pernicious cascade-effects and group polarization, including in business, government, and the judicial system. But the tendency to conform can also lead to *personally* sub-optimal outcomes, as Mill suggests, if it leads people to have stunted characters and miss out on developing their capacities to the highest degree.

Yet conformity and deference are by no means always pernicious. It is worth noting that conformity behavior is often driven by respectable informational needs (Sunstein 2019). Moreover, it should be clear that deference plays a crucial role in the cooperative division of labor. All societies have some practices of expertise, teaching, leadership, and cooperative practices of information sharing (Henrich 2017, Sterelny 2012). Reliance on, and deference to, appropriate

others is indispensable. In short, if tendencies to conformity and deference can be personally and socially sub-optimal, it is also true that they can be personally and socially valuable.

The goal should, therefore, be to achieve an appropriately calibrated independence of mind. One might conform too readily, change one's opinions too quickly under pressure, outsource one's judgment too often, just as one might be overly resistant to pressure, including rational or informationally significant pressure from epistemic peers and betters, refusing to dial down one's confidence or change one's attitudes in response to evidence that one ought to do so.

The value of mental independence is like the value of courage. While the emotion of fear plays an indispensable role in human life, it can also keep us from embarking on valuable courses of action or deter us from staying the course. Because fear can so often win the day, courage is generally a valuable antidote, protecting against our liability to be ruled by fear. Yet courage can be overblown or misplaced. The goal must therefore be to temper courage with normative sensitivity so that one fights fear back only when that is the appropriate thing to do.

Similarly, the importance of an independent mind speaks to our deep intellectual dependence and social influenceability. The human cognitive milieu is profoundly social—a fact that goes deep into our evolutionary past. Humans are primed for conformity and cultural learning, spending much of their prolonged childhoods being socialized into the norms, values, and customs of their environment (Henrich 2017, Sterelny 2012, Tomasello 2019). Moreover, these tendencies toward conformity are coupled with acute sensitivity to power and prestige, tending to make people deferential toward those in power and those with high social status (Haidt & Joseph 2008, Henrich 2017). Our deep intellectual dependence and social influenceability clearly play vital roles in human social life, enabling cumulative cultural learning and facilitating complex cooperation, among other things. But as already noted, the same tendencies can lead to socially and individually

sub-optimal outcomes. Hence, independence of mind can be valuable as a general prophylaxis against problematic forms of conformity and deference. At the same time, if it is not appropriately circumspect, independence of mind can quickly turn into vicious obstinacy. The goal must be to regulate independence of mind so that it reaps cooperative benefits while avoiding social and individual damage.

On the reason-first view, as I explained in chapter 3, the value of mental independence is regulated by the aim of getting it right. I distinguished three different forms of mental independence: responsiveness to reasons *as one sees them*, responsiveness to reasons *as indicated by one's evidence*, and responsiveness to reasons *as they are*. These can be seen as successive enrichments of mental independence. Within the reason-first view, responsiveness to reasons *as they are* is the highest-grade good and makes sense of the value of the others. The kind of independence of mind I have been describing as analogous to courage is on the first rung. Resisting peer pressure, not being too quick to conform, not being overly deferential, and so on—these are most naturally described in terms of responding to reasons as one sees them. The person who responds to reasons as she sees them does not allow her sense of the reasons at issue to be determined or settled by what others think; she has her own take on the reasons. But, ideally, such responsiveness to reasons should be regulated by responsiveness to reasons as actually indicated by one's evidence. And sensitivity to reasons indicated by one's evidence is ultimately subservient to the higher goal of succeeding in tracking fact-relative reasons. That is, we might say, the whole point of following the evidence where it leads.

5. Self-Governance

Self-governance can mean different things. If by "self-governance" one means authenticity, or independence, or responsibility-entailing freedom, or perfection-entailing freedom, then the answer to the question "What is valuable about self-governance?" will coincide with the question of what is valuable about these other items. As I argued in chapter 4, there is a further sense of self-governance which, though very much related to these other things, is not the same as any one of them. As I interpreted it, self-governance is a kind of active self-management in the service of valued goals. My interest in self-governance in this sense is that it is a sort of implementation-mechanism for many of the central goods associated with personal autonomy. For rationally frail and wayward creatures like ourselves, self-governance amounts, in many ways, to a form of active limitation-management, being a vehicle whereby we engage in the process of discerning and conforming to reasons. So conceived, the value of self-governance is largely instrumental. It is a kind of active self-stewardship which brings about other goods valued for their own sake.

6. Autonomy and Dignity

I have been examining the values associated the different strands of our thinking about autonomy. It is fitting to conclude by taking a step back and considering autonomy's value more comprehensively. In an evocative passage, Isaiah Berlin (1958/2002: 178) describes the kind of freedom he wants to enjoy:

> I wish my life and decisions to depend on myself, not on external forces of whatever kind. I wish to be an instrument of my own, not of other men's, acts of will. I wish to be a subject, not an object; to be moved by reasons, by conscious purposes, which are my own, not by causes which affect me, as it were, from outside. I wish to be somebody, not nobody; a doer—deciding, not being decided for, self-directed and not acted upon by external nature or by other men as if I were a thing, or an animal,

> or a slave incapable of playing a human role, that is, of conceiving goals and policies of my own and realizing them. This is at least part of what I mean when I say that I am rational, and that it is my reason that distinguishes me as a human being from the rest of the world. I wish, above all, to be conscious of myself as a thinking, willing, active being, bearing responsibility for my choices and able to explain them by reference to my own ideas and purposes. I feel free to the degree that I believe this to be true, and enslaved to the degree that I am made to realize that it is not.

Although Berlin does not use the language of autonomy, his description of what is desirable about (positive) freedom captures a great deal of what philosophers have thought is desirable about autonomy, and it does so in an intuitively compelling way.

Berlin paints a striking picture here of the thing he wants, apparently for its own sake, namely, to be an agent in the world, and, moreover, an agent of a particular kind. He wants his life and decisions to depend on himself, not on external forces, on conscious purposes which are his own, not on causes which affect him "from the outside," so that ultimately the explanation of his choices must make reference to *him*, to his own ideas and purposes, rather than forces and wills beyond him. Sticks and stones are essentially inert. They are non-agents, physically determined by the laws that operate on them. Explanations of their behavior do not run through anything like a conscious or subjective perspective on the world.

Mere conscious agency won't do either. Slaves are conscious agents, but they are socially dominated and controlled. Their lives and activities are determined by others, and a great deal of their behavior is explained by reference to these others. Animals, too, are conscious agents. They engage the world as purposive beings, and their behavior is explained by states internal to them. But they are, in a sense, slaves to their inner drives and instincts. Explanations of their behavior makes reference to such drives and instincts, not to the kind of reflective and responsible agency we associate with humans.

Berlin wants not just to be a conscious agent, but to be a free, responsible, and self-determining agent. He wants to be moved by his own conscious deliberations and choices for which *he* can, in some sense, be credited, *not* to be pushed around by mere forces, whether natural or social, internal or external. What Berlin is after, then, might be described as the most comprehensive and fundamental agency value: to be an active, responsible, and fully human agent.

This agency value is often associated with dignity. One of the earliest sources of the idea of human dignity connected to free and autonomous agency is the Italian Renaissance humanist, Pico della Mirandola. In his 1496 *Oration on the Dignity of Man*, Mirandola (quoted in Dworkin 1988: 13) has God say to Adam, "We have given thee, Adam, no fixed seat, no form of thy very own, no gift peculiarly thine, that…thou mayest… possess as thine own the seat, the form, the gift which though thyself shalt desire…thou wilt fix the limits of thy nature for thyself…thou…art the molder and the maker of thyself." Two centuries later, Joseph Butler (1726/1983: 15) echoes the idea: "A machine is inanimate and passive, but we are agents. Our constitution is put in our power. We are charged with it; and therefore are accountable for any disorder or violation of it." Implicit in the theological context shared by Mirandola and Butler is the idea that humans are made *Imago Dei*, in the image of God. They are given the gift of free and responsible self-creation, though they are also accountable for their creative activity to the one who has bestowed it. In the course of modern philosophy, the ideas of dignity and autonomy become decoupled from theological commitments (Schneewind 1998). But the idea of self-creation continues to find resonance in secular ethical thought, unhitched from any idea of accountability to God. Jean Paul Sartre (1946/2007) perhaps best epitomizes the idea in his famous phrase "existence precedes essence." Human freedom, for Sartre, is supposed to be grounded in the fact that, unlike other creatures, we

have no antecedent and determinate nature: we can choose what we wish to be, fashion our values, create our sense of meaning and purpose.

The Sartrean model of self-creation is very much alive in formal views of autonomy. For on those views, there is nothing fundamentally for the autonomous will to be beholden to. One must simply decide what one wants to be, who one is, what one stands for. The reason-first approach I have been developing in the course of this dissertation suggests a certain affinity with the earlier views of self-creation. While it gets rid of the idea of theological accountability, it retains the idea of an independent order of normative truths which are *not* our creation and which both constrain and make possible our free and autonomous agency. Ultimately, to be genuine and meaningful self-creators, the view maintains that we must be responsible agents, and this requires us to be undeluded about the world of value in which we live and move. Instead of seeing this as requiring freedom from a fixed human nature, the reason-first view sees autonomy, fundamentally, as a kind of fulfillment of potentialities latent within our nature.

At its heart, then, autonomy is an agency value connected to our dignified status as persons. One of the deepest facts about us human beings is that we are agents, and agents of a certain sort. It is because of the kinds of creatures we are that autonomy is both possible and valuable for us. The reason-first model of autonomy interprets what is special about us in terms of normative agency, and it understands normative agency in a particular way, viz. as defined by an aim-governed conception of practical reason oriented toward genuine normative truths about reasons and values, and by normative capacities which put success at that aim within reach.

Conclusion

A recurrent theme in this dissertation has been that our thinking about autonomy is complex. In briefly exploring the value of each of the conceptual strands associated with our thinking about autonomy, this chapter has attempted to make a similar point about value. Just as there can be no simple answer to the question, "What is autonomy?", there can be no simple answer to the question, "What is the value of autonomy?" But I have also been making the case for an overarching framework that can help us think about what holds our various autonomy concerns together and gives them a unified structure. My goal in this chapter has been to further demonstrate this combination of complexity and unity. While complexity demands that we pay attention to different strands of value and describe them with appropriate texture and nuance, the framework I have been developing also supplies a kind of master value in terms of which we can think fruitfully about why autonomy matters.

References

Ainslie, G. (1974). "Impulse Control in Pigeons 1." *Journal of the Experimental Analysis of Behavior*, *21*(3), 485-489.

Arneson, R. (1980). "Mill Versus Paternalism." *Ethics*, *90*(4), 470-489.

Arpaly, N. (2002). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford University Press.

——— . (2002). "Moral Worth." *The Journal of Philosophy*, *99*(5), 223-245.

——— . (2005). "Responsibility, Applied Ethics, and Complex Autonomy Theories." In Taylor (ed.), *Personal autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge University Press.

Asch, S. (1955). "Opinions and Social Pressure." *Scientific American*, *193*(5), 31-35.

Bagley, B. (2013). *Improvisational Agency* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Bargh, J. (2017). *Before You Know It: The Unconscious Reasons We Do What We Do*. Simon and Schuster.

Baumeister, R., Vohs, K., & Tice, D. (2007). "The Strength Model of Self-control." *Current Directions in Psychological Science*, *16*(6), 351-355.

Beauchamp, T. (2005). "Who Deserves Autonomy and Whose Autonomy Deserves Respect." In Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge University Press.

Beauchamp, T., & Childress, J. (2008). *Principles of Biomedical Ethics* (6th edition), Oxford: Oxford University Press.

Benn, S. (1975). "Freedom, Autonomy and the Concept of a Person." *Proceedings of the Aristotelian Society* 76, 109-130.

Benson, P. (1987). "Freedom and Value." *The Journal of Philosophy*, *84*(9), 465-486.

——— . (1990). "Feminist Second Thoughts about Free Agency." *Hypatia*, *5*(3), 47-64.

——— . (1994). "Free Agency and Self-Worth." *The Journal of Philosophy*, *91*(12), 650-668.

——— . (2005). "Feminist Intuitions and the Normative Substance of Autonomy." In Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge University Press.

Beran, M. (2015). "Chimpanzee Cognitive Control." *Current Directions in Psychological Science*, 24(5), 352-357.

Berlin, I. (1958/2002). "Two Concepts of Liberty." In Hardy, H. (ed.), *Liberty*. Oxford University Press.

Berofsky, B. (1995). *Liberation from Self: A theory of Personal Autonomy*. Cambridge University Press.

Bratman, M. (1999). "Intention, Decision, and Treating as a Reason." In *Faces of Intention: Selected Essays on Intention and Agency*, Cambridge University Press.

——— . (2003). "Autonomy and Hierarchy." *Social Philosophy and Policy*, *20*(2), 156-176.

——— . (2004). "Three Theories of Self-governance." *Philosophical Topics*, *32*(1/2), 21-46.

——— . (2009). "Intention, Practical Rationality, and Self-Governance." *Ethics*, *119*(3), 411-443.

——— . (2018). *Planning, Time, and Self-Governance: Essays in Practical Rationality*. Oxford University Press.

Braunstein, L., Gross, J., & Ochsner, K. (2017). "Explicit and Implicit Emotion Regulation: A Multi-level Framework." *Social Cognitive and Affective Neuroscience*, *12*(10), 1545-1557.

Brink, D. (1989). *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

——— . (2007). *Perfectionism and the Common Good: Themes in the Philosophy of T.H. Green*. Clarendon Press.

——— . (2008). "The Significance of Desire." *Oxford Studies in Metaethics*, *3*, 5-45.

——— . (2013). *Mill's Progressive Principles*. Oxford University Press.

——— . (2013). "Situationism, Responsibility, and Fair Opportunity." *Social Philosophy and Policy*, *30*(1-2), 121-149.

——— . (2019). "Normative Perfectionism and the Kantian Tradition." Philosophers' Imprint 19(45).

Brink, D., & Nelkin, D. (2013). "Fairness and the Architecture of Responsibility." In Shoemaker, D (ed.), *Oxford Studies in Agency and Responsibility*, *1*.

Bublitz, J., & Merkel, R. (2009). "Autonomy and Authenticity of Enhanced Personality Traits." *Bioethics*, *23*(6), 360-374.

Burger, J. (2009). "Replicating Milgram: Would People Still Obey Today?" *American Psychologist*, *64*(1), 1.

Buss, S. (2012). "Autonomous Action: Self-determination in the Passive Mode." *Ethics*, *122*(4), 647-691.

Butler, J. (1983). *Five Sermons*. Darwall, S. (ed.). Hackett Publishing.

Carter, I. (2011). "Respect and the Basis of Equality." *Ethics*, *121*(3), 538-571.

Chang, R. (2000). "Voluntarist Reasons and the Sources of Normativity." In Sobel, D. & Wall, S. (eds), *Reasons for Action*. Cambridge University Press.

Christman, J. (1988). "Constructing the Inner Citadel: Recent Work on the Concept of Autonomy." *Ethics*, *99*(1), 109-124.

——— . (1991a). "Autonomy and Personal History." *Canadian Journal of Philosophy*, *21*(1), 1-24.

——— . (1991b). "Liberalism and Individual Positive Freedom." *Ethics*, *101*(2), 343-359.

——— . (2005). "Procedural Autonomy and Liberal Legitimacy." In Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge University Press.

——— . (2009). *The Politics of Persons: Individual Autonomy and Socio-Historical Selves*. Cambridge University Press.

——— . (2018). "Autonomy in Moral and Political Philosophy," *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2018/entries/autonomy-moral/>.

Cohen, J. (2017). "The Basics of Cognitive Control: Theoretical Constructs and Behavioral Phenomena." In Egner, T. (ed.), *The Wiley Handbook of Cognitive Control*. John Wiley & Sons.

Collier, P., & Betts, A. (2017). *Refuge: Rethinking Refugee Policy in a Changing World*. Oxford University Press.

Cooper, J. (2003). "Stoic Autonomy." In Paul, E., Miller, F., & Paul, J. (eds.), *Autonomy: Volume 20*. Cambridge University Press.

Dalley, J., Cardinal, R., & Robbins, T. (2004). "Prefrontal Executive and Cognitive Functions in Rodents: Neural and Neurochemical Substrates." *Neuroscience & Biobehavioral Reviews*, *28*(7), 771-784.

Darley, J. & Batson, C. (1973). "From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior." *Journal of personality and social psychology*, *27*(1), 100.

Darwall, S. (1977). "Two Kinds of Respect." *Ethics*, *88*(1), 36-49.

——— . (2006). "The Value of Autonomy and Autonomy of the Will." *Ethics*, *116*(2), 263-284.

Doris, J. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Dworkin, G. (1988). *The Theory and Practice of Autonomy*. Cambridge University Press.

Dworkin, R. (1993). *Life's Dominion: An Argument about Abortion, Euthanasia, and Individual Freedom*. Vintage Press.

Ekstrom, L. (2005). "Autonomy and Personal Integration." In Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge University Press.

Elster, J. (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality.* Cambridge University Press.

——— . (2000). *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*. Cambridge University Press.

Enoch, D. (2017). "Hypothetical Consent and the Value(s) of Autonomy." *Ethics*, *128*(1), 6-36.

Epictetus. (2018). *How to be Free: An Ancient Guide to the Stoic Life* (enchiridion and discourses). Tansl. by A.A. Long. Princeton University Press.

Feinberg, J. (1986). *The Moral Limits of the Criminal Law*. Vol. III, Harm to Self. Oxford University Press.

Feldman, F. (2004). *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford University Press on Demand.

Figueira, T. (1990). "Autonomoi kata tas spondas" (Thucydides 1.67. 2). *Bulletin of the Institute of Classical Studies*, (37), 63-88.

Fischer, J. & Ravizza, M. (1998). *Responsibility and Control*. Cambridge University Press.

Frankfurt, H. (1971). "Freedom and the Concept of a Person." Journal of Philosophy. 68, 829-839.

——— . (1988). *The Importance of What We Care About*. Cambridge University Press.

—— . (1999). “Autonomy, Necessity, and Love.” In *Necessity, Volition, and Love.* Cambridge University Press.

—— . (2006). *Taking Ourselves Seriously and Getting it Right*. Stanford University Press.

Friedman, M. (2003). *Autonomy, Gender, Politics*. Oxford University Press.

Garnett, M. (2013). “Taking the Self out of Self-rule.” *Ethical Theory and Moral Practice*, *16*(1), 21-33.

—— . (2014). “The Autonomous Life: A Pure Social View.” *Australasian Journal of Philosophy*, *92*(1), 143-158.

Gewirth, A. (1996). *The Community of Rights*. University of Chicago Press.

Gollwitzer, P. (1999). “Implementation Intentions: Strong Effects of Simple Plans.” *American psychologist*, *54*(7), 493.

Graham, P. (2010). “In Defense of Objectivism about Moral Obligation.” *Ethics*, *121*(1), 88-115.

Green, T. H. (1886/1986). “On the Different Senses of ‘Freedom’ as applied to Will and to the Moral Progress of Man.” In Harris, P. & Morrow, J. (eds.), *Lectures on the Principles of Political Obligation and Other Writings*, Cambridge University Press.

Griffin, J. (2008). *On Human Rights*. Oxford University Press.

Groll, D. (2012). “Paternalism, Respect, and the Will.” *Ethics*, *122*(4), 692-720.

Gross, J. (ed.). (2014). *Handbook of Emotion Regulation*. Guilford Publications.

—— . (2015). “Emotion Regulation: Current Status and Future Prospects.” *Psychological Inquiry*, *26*(1), 1-26.

Haidt, J., & Joseph, C. (2008). “The Moral Mind: How Five Sets of Innate Intuitions Guide the Development of Many Culture-specific Virtues, and perhaps even Modules.” In Carruthers, P., Laurence, S., & Stich, S. (eds), *The Innate Mind* (Vol. 3). Oxford University Press.

Henrich, J. (2017). *The Secret of our Success: How Culture is Driving Human Evolution, Domesticating our Species, and Making Us Smarter*. Princeton University Press.

Hill, T. (1991). *Autonomy and Self-Respect*. Cambridge University Press.

Hofmann, S., Asnaani, A., Vonk, I., Sawyer, A., & Fang, A. (2012). “The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses.” *Cognitive therapy and research*, *36*(5), 427-440.

Hurka, T. (1987). "Why Value Autonomy?" *Social Theory and Practice*, *13*(3), 361-382.

—— . (1993). *Perfectionism*. Oxford University Press.

Isen, A. & Levin, P. (1972). "Effect of Feeling Good on Helping: Cookies and Kindness." *Journal of Personality and Social Psychology*, *21*(3), 384.

Ismael, J. (2016). *How Physics Makes Us Free*. Oxford University Press.

Jaworska, A. (2007). "Caring and Internality." *Philosophy and Phenomenological Research*, *74*(3), 529-568.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Macmillan Publishers.

Kane, R. (1998). *The Significance of Free Will*. Oxford University Press.

Kant, I. (1999). *Practical Philosophy*. Gregor, M. (ed). Cambridge University Press.

Kauppinen, A. (2011). "The Social Dimension of Autonomy." In Petherbridge (ed.), *Axel Honneth: Critical Essays*. Brill.

Killmister, S. (2013). "Autonomy and False Beliefs." *Philosophical Studies*, *164*(2), 513-531.

—— . (2018). *Taking the Measure of Autonomy*. Routledge.

Korsgaard, C. (1996). *The Sources of Normativity*. Cambridge University Press.

—— . (2009). *Self-Constitution*. Oxford University Press.

Kross, E., & Ayduk, O. (2011). "Making Meaning out of Negative Experiences by Self-Distancing." *Current Directions in Psychological Science*, *20*(3), 187-191.

Long, A. (2015). *Greek Models of Mind and Self* (Vol. 22). Harvard University Press.

Mackenzie, C. (2014). "Three Dimensions of Autonomy." In Veltman, A., & Piper, M. (eds), *Autonomy, Oppression, and Gender*. Oxford University Press.

MacLean, E., Hare, B., Nunn, C., Addessi, E., Amici, F., Anderson, R., ... & Boogert, N. (2014). "The Evolution of Self-control." *Proceedings of the National Academy of Sciences*, *111*(20), E2140-E2148.

Maclure, J., & Taylor, C. (2011). *Secularism and Freedom of Conscience*. Harvard University Press.

McDowell, J. (2010). “Autonomy and its Burdens.” *The Harvard Review of Philosophy*, *17*(1), 4-15.

McHugh, C. (2017). “Attitudinal Control.” *Synthese*, *194*(8), 2745-2762.

McKenna, M. (2005). “The Relationship Between Autonomous and Morally Responsible Agency.” In Taylor (ed.), *Personal Autonomy: New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*. Cambridge University Press.

Mele, A. (1995). *Autonomous Agents: From Self Control to Autonomy*. Oxford University Press.

—— . (2014). *Free: Why Science Hasn't Disproved Free Will*. Oxford University Press.

Mercier, H., & Sperber, D. (2017). *The Enigma of Reason*. Harvard University Press.

Meyers, D. (2004). *Being Yourself: Essays on Identity, Action, and Social Life*. Rowman & Littlefield Publishers.

—— . (2005). “Decentralizing Autonomy: Five Faces of Selfhood.” In Anderson and Christman (eds.), *Autonomy and the Challenges to Liberalism*. Cambridge University Press.

Milgram, S. (1974/2009). *Obedience to Authority*. Harper Perennial Modern Classics, Reprint Edition.

Mill, J.S. (1859/2003). *On Liberty.* Kateb and Bromwich (eds.). Yale University Press.

Mischel, W. (2014). *The Marshmallow Test*. Bantam Press.

Mischel, W., Shoda, Y., & Rodriguez, M. (1989). “Delay of Gratification in Children.” *Science*, *244*(4907), 933-938.

Möller, K. (2012). *The Global Model of Constitutional Rights*. Oxford University Press.

Montero, B. (2016). *Thought in Action: Expertise and the Conscious Mind*. Oxford University Press.

Nagel, T. (1989). *The View from Nowhere*. Oxford University Press.

Neal, D., Wood, W., & Quinn, J. (2006). “Habits—A repeat Performance.” *Current Directions in Psychological Science*, *15*(4), 198-202.

Nelkin, D. (2005). “Freedom, Responsibility and the Challenge of Situationism.” *Midwest Studies in Philosophy*, 29.

—— . (2011). *Making Sense of Freedom and Responsibility*. Oxford University Press.

—— . (2016). “Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness.” *Nous*, 50(2), 356-378.

—— . (forthcoming). “Free Will and Aesthetic Responsibility.” In Uidhir, C. (Ed.), *Art and Philosophy*. Oxford University Press.

Noggle, R. (1995). “Autonomy, Value, and Conditioned Desire.” *American Philosophical Quarterly*, *32*(1), 57-69.

Nussbaum, M. (2011). *Creating Capabilities*. Harvard University Press.

Oshana, M. (1998). “Personal Autonomy and Society.” *Journal of Social Philosophy*, 29(1): 81-102.

—— . (2006). *Personal Autonomy in Society*. Ashgate Publishing.

Parfit, D. (2011). *On What Matters* (Vol. 1). Oxford University Press.

Pennebaker, J. W. (1997). “Writing about Emotional Experiences as a Therapeutic Process.” *Psychological science*, *8*(3), 162-166.

Pereboom, D. (2014). *Free Will, Agency, and Meaning in Life*. Oxford University Press.

Pettit, P. (1997). *Republicanism: A Theory of Freedom and Government*. Oxford University Press.

—— . (2001). *A theory of Freedom: From the Psychology to the Politics of Agency*. Oxford University Press.

—— . (2014). *Just Freedom: A Moral Compass for a Complex World*. WW Norton & Company.

Pettit, P., & Smith, M. (1990). “Backgrounding Desire.” *The Philosophical Review*, *99*(4), 565-592.

—— . (1993). “Practical Unreason.” *Mind*, *102*(405), 53-79.

—— . (1996). “Freedom in Belief and Desire.” *The Journal of Philosophy*, *93*(9), 429-449

Plato. (1997). *Complete Works*. John Cooper (ed). Hackett Publishing.

Putnam, H. (1975). *Mind, Language, and Reality* (vol. 2). Cambridge University Press.

Railton, P. (2006). “Normative Guidance.” In Shafer-Laundau, R. (ed), *Oxford Studies in Metaethics*, *1*, 3-34.

—— . (2009). “Practical Competence and Fluent Agency.” In Sobel, D., & Wall, S. (eds), *Reasons for Action*. Cambridge University Press, 81-115.

Rawls, J. (1971/1999). *A theory of Justice*. Harvard university press.

Raz, J. (1986). *The Morality of Freedom*. Oxford University Press.

—— . (1997). "The Active and the Passive." *Proceedings of the Aristotelian Society, Supplementary Volumes*, *71*, 211-246.

—— . (1999). *Engaging Reason: On the Theory of Value and Action*. Oxford University Press.

Rosen, G. (2003). "Culpability and Ignorance." *Proceedings of the Aristotelian Society* 103(1), 61-84.

Roskies, A. (2012). "Don't Panic: Self-Authorship without Obscure Metaphysics." *Philosophical Perspectives*, *26*(1), 323-342.

Ross, L., & Nisbett, R. (2011). *The Person and the Situation: Perspectives of Social Psychology*. Pinter & Martin Publishers.

Sartre, J. (1946/2007*). Existentialism is a Humanism*. Yale University Press.

Savulescu, J. (1995). "Rational Non-interventional Paternalism: Why Doctors Ought to Make Judgments of What is Best for their Patients." *Journal of Medical Ethics*, *21*(6), 327-331.

Sayre-McCord, J. & Smith, M. (2014). "Desires…and Beliefs…of One's Own." In Vargas, M., & Yaffe, G. (eds). *Rational and Social Agency: The Philosophy of Michael Bratman*. Oxford University Press.

Scanlon, T. (1972). "A Theory of Freedom of Expression." Philosophy & Public Affairs *204*(2), I5-20.

—— . (1998). *What we Owe to each Other*. Harvard University Press.

Scheffler, S. (1994). *The Rejection of Consequentialism.* Oxford University Press.

—— . (2011). "Valuing." In Wallace, R. J., Kumar, R., & Freeman, S. (eds). *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*. Oxford University Press.

Schelling, T. (1978). "Egonomics, or the Art of Self-management." *The American Economic Review*, *68*(2), 290-294.

Schneewind, J. (1998). *The Invention of Autonomy.* Cambridge University Press.

Sen, A. (1999). *Development as Freedom.* Oxford University Press.

Shepherd, J. (2014). "The Contours of Control." *Philosophical Studies*, *170*(3), 395-411.

Sher, G. (1997). *Beyond Neutrality: Perfectionism and Politics*. Cambridge University Press.

Sherif, M. (1965). *The Psychology of Social Norms*. Octagon Books.

Sobel, D. (2009). "Subjectivism and Idealization." *Ethics*, *119*(2), 336-352.

Sripada, C. (2016). "Self-expression: A Deep Self Theory of Moral Responsibility." *Philosophical Studies*, *173*(5), 1203-1232.

—— . (manuscript). "At the Center of Agency, The Deep Self"

Sterelny, K. (2012). *The Evolved Apprentice*. MIT press.

Stoljar, N. (2000). "Autonomy and the Feminist Intuition." In Mackenzie and Stoljar (eds.), *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*. Oxford University Press.

—— . (2014). "Autonomy and Adaptive Preference Formation." In Veltman, A., & Piper, M. (Eds.). *Autonomy, Oppression, and Gender*. Oxford University Press.

Suhler, C., & Churchland, P. (2009). "Control: Conscious and Otherwise." *Trends in Cognitive Sciences*, *13*(8), 341-347.

Terlazzo, R. (2016). "Conceptualizing Adaptive Preferences Respectfully: An Indirectly Substantive Account." *Journal of Political Philosophy* 24(2).

Thaler, R., & Sunstein, C. (2009). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Penguin.

Theunissen, N. (2018). "Must We Be Just Plain Good? On Regress Arguments for the Value of Humanity." *Ethics*, *128*(2), 346-372.

Tierney, B. (2014). *Liberty and Law* (Vol. 12). Catholic University Press.

Tomasello, M. (2019). *Becoming Human: A theory of Ontogeny*. Belknap Press.

Vargas, M. (2006). "Review of Personal Autonomy: New Essays on Personal Autonomy and Its Role in Contemporary Morality." *Notre Dame Philosophical Reviews* (8/15).

—— . (2013a). *Building Better Beings: A Theory of Moral Responsibility*. Oxford University Press.

—— . (2013b). "Situationism and Moral Responsibility: Free Will in Fragments." Clark, A., Kiverstein, J., Vierkant, T. (eds), *Decomposing the Will*. Oxford University Press.

——— . (2017). “Implicit Bias, responsibility, and Moral Ecology.” In Shoemaker, D. (ed.), *Oxford Studies in Agency and Responsibility*, *4*, 219-247.

——— . (2018). “The Social Constitution of Agency and Responsibility: Oppression, Politics, and Moral Ecology.” In Hutchison, K., Mackenzie, C., & Oshana, M. (eds), *The Social Dimensions of Responsibility*. Oxford University Press.

Velleman, J. (2000). *The Possibility of Practical Reason*. Oxford University Press.

Vierkant, T. (2013). “Managerial Control and Free Mental Agency.” In Clark, A., Kiverstein, J., Vierkant, T. (eds), *Decomposing the Will*. Oxford University Press.

Waldron, J. (2017). *One Another’s Equals: The Basis of Human Equality*. Harvard University Press.

Wall, S. (1998). *Liberalism, Perfectionism and Restraint*. Cambridge University Press.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. Harvard University Press.

——— . (2014). “Reasons, Policies, and the Real Self: Bratman on Identification.” In Vargas, M., & Yaffe, G. (Eds.). *Rational and Social Agency: The Philosophy of Michael Bratman*. Oxford University Press, USA.

Watson, G. (1975). “Free Agency.” *The Journal of Philosophy*, *72*(8), 205-220.

——— . (1996). “Two Faces of Responsibility.” *Philosophical Topics*, *24*(2), 227-248.

Westlund, A. (2003). “Selflessness and Responsibility for Self: Is Deference Compatible with Autonomy?” *The Philosophical Review*, *112*(4), 483-523.

——— . (2009). “Rethinking Relational Autonomy.” *Hypatia*, *24*(4), 26-49.

Wilson, T. (2004). *Strangers to Ourselves*. Harvard University Press.

——— . (2011). *Redirect: The Surprising New Science of Psychological Change*. Penguin.

Wolf, S. (1994). *Freedom Within Reason*. Oxford University Press.

Wolff, R. (1970/1998). *In Defense of Anarchism*. University of California Press.

Zimbardo, P., Haney, C., Banks, W., & Jaffe, D. (1973). “A Pirandellian Prison: The Mind is a Formidable Jailer.” *New York Times Magazine*, *8*(1973), 38-60.

Zimmerman, M. (2014). *Ignorance and Moral Obligation*. Oxford University Press.