London School of Economics and Political Science

Doctoral Thesis

Measuring Freedom, and its Value

Author:
Nicolas Cote

Supervisor:
Prof. Alex Voorhoeve
Dr. Campbell Brown

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Department of Philosophy, Logic, and the Scientific Method

October 24, 2021

Declaration of Authorship

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others.

I confirm that a version of chapter 2 is published as "Weakness of Will and the Measurement of Freedom" in *Ethics* 130 (3): 384-414.

I declare that this thesis consists of 86,600 words.

Abstract

This thesis concerns the measurement of freedom, and its value. Specifically, I am concerned with three overarching questions. First, can we measure the extent of an individual's freedom? It had better be that we can, otherwise much ordinary and intuitive talk that we would like to vindicate – say, about free persons being freer than slaves – will turn out to be false or meaningless. Second, in what ways is freedom valuable, and how is this value measured? It matters, for example, whether freedom is valuable only insofar as it enables us to pursue specific ends we happen to value for independent reasons, or whether it is also valuable in itself. In the latter case, but not the former, more freedom will always (other things being equal) be better. Likewise, it's important to get clarity on what ends, exactly, freedom is especially instrumentally valuable in promoting, since this goes to how much we should care about it. And finally, the liberal political tradition asserts that there is a special sphere of personal choices within which individuals should be free to do as they please, for reasons over and above the value of freedom itself. Now, can we measure the extent to which states and individuals respect individual liberties, and can we weigh the importance of respecting liberty against competing values like social welfare?

To answer these questions, I follow the axiomatic tradition of social choice theory, and I develop several novel measures of freedom – including the first measure of freedom ever proposed that is sensitive to how modally robustly our options are available – a novel approach to measuring the diversity of an opportunity set, and I develop an account of the value of freedom according to which an important part of why freedom matters is that it enables us to improve our preferences through learning, and I construct ways of measuring the value of freedom along this dimension of value. Finally, in later chapters, I provide a representation result for a measure of illiberalism, i.e. of the degree to which states fail to respect the rights of their citizens, and I extend this result to provide a characterization of moral theories that express concern for respecting the rights of others. The final chapter closes by discussing how this concern may be weighed against concerns for social welfare.

Acknowledgements

My deepest gratitude goes out to my supervisors Alex Voorhoeve and Campbell Brown. I am deeply thankful for all the advice and encouragement they gave me, for the hours of discussion we have shared, and for all that I learned while working under their guidance. Their feedback was always incisive, probing, and supportive. Much in the way that John Lennon and Paul McCartney never walked away from a songwriting session without having something to show for it, I never walked away from a supervision with either Alex or Campbell without having much material with which to improve my research. I can only hope that my work in this thesis will do them proud.

The LSE's philosophy department has been a wonderful and stimulating environment for me in these past four years. I never felt more at home in a philosophy department, and I'm terribly sorry that my time here is at an end. The faculty were incredibly generous with their time for me, both as regards my thesis work and my non-thesis work, and I enjoyed the friendship of a terrific crowd of PhD students, whose warm company has made these years so enjoyable and rewarding. So, my thanks to Richard Bradley, who provided valuable feedback on a number of my chapters, to Johanna Thoma, Susanne Burri, and Liam Kofi-Bright, who helped me in other work, to Mike Otsuka, who greatly improved my quality of life in London by so kindly treating me to *so many* unforgettable opera performances at the ROH, and to my dear friends in the PhD program, Bastian, Chloé, Ko-Hung, Joe, James, and Margherita.

I am also indebted to Philip Pettit. Though he did not supervise me in my thesis work, he hosted me at the ANU during the second half of my third year, and I am deeply grateful for all the hours of discussion on my work we shared there, for his many pointed objections, and for the warmth and gregariousness with which he welcomed me. Though my time there was cut short by the pandemic, I will always remember it fondly.

Many thanks also to Nicholas Baigent, John Weymark, Alan Hajèk, audiences at the LSE's Choice Group, the 2019 formal ethics conference in Ghent, the participants in a recurring zoominar organized by Hendrik Rommeswinkel on freedom of choice, and to the audience of a informal Euro-Australian zoominar I organized on republican freedom.

Finally, my most special thanks to my parents, Victoria and Antoine. For their love and support, for all the comments they gave me on my work, for all their advice and encouragement, for absolutely everything that they did.

Contents

Introduction

Canada is free and freedom is its Nationality ~ Wilfrid Laurier

I am a Canadian, a free Canadian... This heritage of freedom I pledge to uphold for myself and all mankind. ~ John Diefenbaker

We peer so suspiciously at one another that we cannot see that we Canadians are standing on the mountaintop of human wealth, freedom, and priviledge. ~ Pierre Trudeau

WHatever positive associations come to mind when you think of Canada – maple syrup, beaver tails, hockey, politeness, Inuit throat singing, Ryan Gosling – freedom is unlikely to feature among them. Perhaps this is not so surprising. After all, our revolt against the British failed, we have a sordid history of repression against indigenous peoples, and we have to compete for attention on the world stage with our neighbours down south, who are much noisier than we are in boasting about how much they love freedom. Still, it's not for lack of good intentions on our part. A look at the history of Canadian political thought reveals some of the most prolific writing in defense of multiculturalism, the open society, open borders, and a surprising number of libertarians (left and right). Our most eminent socialist and communitarian thinkers (G.A. Cohen, Charles Taylor) write as much on freedom as our most notorious anarcho-capitalists (e.g. Jan Narveson). And our former prime ministers certainly enjoyed waxing poetic about it.

It is thus coming on the heels of an established Canadian tradition that I write this thesis, on the subject of freedom. In contrast to my many forebears, though, the questions I take up here are primarily axiological in nature, rather than political. That is, I won't attempt (directly, at least) to slice any of the traditional Gordian knots regarding how far protections on free speech should extend, or how large government should be, etc. Because although freedom is one of the most powerful values in morals and politics, the way we deploy it in ordinary discourse to criticise or justify norms, powers, privileges, and social arrangements rests on the often unargued presupposition that we can talk meaningfully about which arrangements curtail or expand individual freedom the "most," and that there are principled, non-capricious ways of weighing the importance of protecting freedom against other important values and ends. These presuppositions must be defended to secure the well-foundedness of our disputes. Defending them, however, requires us to answer some fundamental questions about the structure and nature of value, and these are the questions at the heart of this thesis.

Specifically, I am concerned with three overarching questions. First, and most fundamentally, can we measure the extent of an individual's freedom? By this I mean: can we design a procedure for assigning real numbers to individuals which represent the individuals' degrees of freedom, such that some important relations between those numbers – notably their order, difference, and ratio – can be meaningfully taken to encode analogous relation between individuals' degrees of freedom? It had better be that we *can* measure freedom, otherwise much ordinary and intuitive talk that we would like to vindicate – say, about free persons being freer than slaves, or about Canadians being on the mountaintop of human freedom – will turn out to be false or meaningless. Second, in what ways is freedom valuable, and how can this value be measured? It matters, for example, whether freedom is valuable only insofar as it enables us to pursue specific ends we happen to value for independent reasons, or whether it is also valuable as an end in itself. In the latter case, but not the former, more freedom will always (other things being equal) be better. Likewise, it's important to get clarity on what ends freedom is especially instrumentally valuable in promoting, since this goes to how much we should care about it. And finally, the liberal political tradition asserts that there is a special sphere of personal choices within which individuals *should* be free to do as they please, for reasons over and above the value of freedom itself. Bader (2018) and R. Dworkin (2011) refer to these specially protected freedoms as "liberties," a convention I will adopt. My third question will therefore be: can we measure the extent to which states, individuals, and other agents respect individual liberties, and can we weigh the importance of respecting liberty against competing values like social welfare? This question is especially relevant for philosophical traditions which hold that people have certain *rights* not to be interfered with, rights which follow from duties of respect, and therefore exert independent pressure on us not to curtail their freedom in certain ways.

These three questions pick up on different aspects of what we care about in relation to freedom. The first two questions concern what Amartya Sen (1993b) has called the "opportunity aspect" of freedom, which is concerned with our actual capability to achieve things that we can and do value, while the third question concerns rather the "process aspect," which has to do with the procedure of free choice itself and the importance of not being interfered with in the exercise of one's free will. As Sen (1993b, p. 522) rightly observes, "a[ny] comprehensive assessment of freedom must take note of both these aspects and the irreducible importance of each of these respective features." It is important to treat all three questions if we want to arrive at a full picture of what freedom is and why it matters. These questions are not simply unified by a common theme, however: they are also unified by a common methodological approach I take to answering them. This approach is characterized by a reliance on the formal tools and methods of social choice and rational choice theory to frame and analyse problems. This is not an innocent methodological choice on my part. Formal tools involve simplifying or idealizing assumptions about agents and the world, which may obscure real difficulties. And of course it's important not to get carried away with formalism to the point of losing contact with the real philosophical problems that motivated us in the first place. But formal models also allow us to state our questions and assumptions very precisely, and when we study the implications of precisely stated assumptions, we often reach surprising insights that can inform philosophical reflection in fruitful ways.

It's worth emphasizing at the outset that, when we talk about measuring some moral good, or the weight of some moral consideration, there are two kinds of questions that might be at issue: one semantic, the other epistemological. When we are disagreeing over how to measure freedom, this might suggest an epistemological disagreement about how to measure an independently specified thing – freedom – whose interpretation we all agree on. If this were so, measuring freedom would be a bit like measuring inflation: there is a clear concept at stake (how much prices rise) but because not all prices rise at the same rate and different people buy different baskets of good, there is no unique rate at which all prices rise nor even an average rate common to all baskets of goods. So economists disagree over how best to measure the overall rate of increase. My own emphasis on formal methods might suggest that this is precisely the sort of problem I am dealing with. But in fact when we're disagreeing over how to measure freedom, the disagreement is really conceptual: as we will see, disagreements over how to measure freedom are really disagreements over what freedom is, and what it is to be freer than another. Philosophers actively disagree over how to interpret the concept of freedom. And many different factors seem to matter to our intuitive judgements, but it's not clear how to combine those concerns into a coherent overall view. Formal modelling is simply a tool for investigating how these concerns can be made to interact.

My thesis is divided into seven chapters. Chapter 1 introduces the problem of the measurement of freedom, and contrasts two different measurement approaches: the axiomatic approach adopted by social choice theorists, and a more informal approach historically favoured by philosophers. Both approaches, thankfully, promise to vindicate the possibility of measuring freedom, but I believe that the formal axiomatic approach is more attractive, as we will see, and so I devote my energies in this chapter to defending it. The most eloquent (and certainly the most cited) defender of the more informal approach is Ian Carter, so he will serve as my foil.

With this argument in place, in chapters 2 and 3 I work within the axiomatic tradition to construct new measures of freedom. Specifically, my aim in chapter 2 is to design measures of freedom that are sensitive to the fact that not only may more or fewer opportunities be available to individuals, but any given opportunity may be more or less *accessible* them. Roughly, the intuition is that (paraphrasing Berlin, 1969) your degree of freedom does not depend merely on how many possibilities are open to you, but also how capable you are of actualizing those possibilities. In chapter 3, in contrast, I aim to construct measures of freedom that are sensitive to a different set of considerations, namely, how many opportunities are available to you and how *robustly* available they are. Numerous philosophers have argued that being free doesn't require merely that others don't prevent you from doing what you will, it requires additionally that no one *could* arbitrarily prevent you from doing what you will. If I could easily, and with impunity, prevent you from doing anything I didn't want you to do, then whatever you *do* do, you do only with my (implicit) permission – you are a subject of my will, not a free person. Constructing a measure of freedom that is sensitive to these modal considerations will require first constructing a measure of "distance" across modal space – i.e. a device that can tell us how much would need to change about the world for some opportunity that is actually available to you to become unavailable.

I transition from the measurement of freedom to the measurement of its value in chapter 4. The motivating question of this chapter is whether the quality of one's opportunities affects how much freedom they offer us, to which my answer is no: the quality of one's opportunities rather affects the *non-specific value* of one's freedom, that is, the value that freedom has in itself, over and above its *specific* value in enabling us to pursue ends we value for independent reasons that have nothing to do with freedom. In taking this line, I commit myself to the view that freedom *has* value in itself, and this is a controversial view, so much of the chapter is devoted to defending this thesis. Of course, the specific value of freedom is also very important, and so I devote chapter 5 to its exploration. The main thesis of this chapter is that there is an important dimension of freedom's specific value which has been overlooked by theorists of freedom, namely, what I call its *diagnostic value*. This is the value it has in providing us with a kind of self-knowledge that improves our abilities as evaluators and choosers. The bulk of this chapter is devoted to discussing the many dimensions of the diagnostic value of opportunities, and along the way I propose ways of measuring diagnostic value along each of its dimensions.

Finally, in chapters 6 and 7, I turn to the discussion of liberty, and of the rights that protect it. In chapter 6 I defend a new formal framework within which to think about rights, and I then operationalize this framework and show how to construct from it several measures of how well the liberty of everyone in society is respected – i.e. a measure that can tell us, say, whether the liberty of residents in Canada is better safeguarded than the liberty of residents in France. An important result of this chapter is that it provides the first formal characterization of how rights constrain social choice that really captures standard philosophical intuitions about the functioning of rights. In chapter 7, I study how the concern for protecting rights interacts with more welfarist concerns. Some philosophers argue that rights have an absolute priority over other, "merely" axiological considerations: you cannot infringe people's rights, no matter the consequences. I argue, using the vocabulary established in chapter 6, that this view cannot be coherently extended to the context of uncertainty without undermining its own motivations. I thus defend a more moderate picture of how concerns for liberty and concerns for welfare interact, providing a simple representation of how individuals who are suitably responsive to both concerns should act.

Chapter 1

Measuring Freedom in Economics and Philosophy

We often speak as though freedom is measurable. I have a firm conviction that I am freer than the typical North Korean, and that ordinary citizens are a great deal freer than prisoners. Mill (1885, p. 81) asserted that the freest country in his day was the United States. Hayek (1960) and Friedman (1962) built their argument for laissez-faire capitalism on the back of the claim that it is the economic system that maximizes individual freedom, a claim which G. A. Cohen (2011) rebuts by arguing that it's really socialism that maximises human freedom. Such claims are rhetorically powerful, but whether they are meaningful depends on the answer to several further questions.

What do we mean when we say of an individual that they are free in some respect? Is it possible to determine which of two individuals enjoys more freedom, given what each of them can do? To answer these questions is to take a stance on whether and to what extent and in what way freedom (in its opportunity aspect) can be measured, and it is the central question that will concern us for the next four chapters.

The measurement question has attracted the attention of both philosophers and welfare economists. Unsurprisingly, different approaches to the measurement of freedom have emerged in the two fields, and they have generated different answers to the above questions. Welfare economists have framed the problem as one of which axiomatic constraints to impose on a rule for ranking finite sets of options in terms of how much freedom they offer. Disagreement in this field thus tends to be marked by disagreement over which axioms are most compelling. In contrast, many philosophers have seen the problem as one of measuring some natural quantity in the world, and here metaphysics has played a big role in defining different authors' positions. For instance, R. Dworkin (1977, p. 172) argues that freedom cannot be measured because freedom is not some "commodity" or physical magnitude which individuals can have in greater or lesser quantities. Likewise, Kristjánsson (1996,

p. 11) claims that freedom cannot be measured because "freedom is a relation, not a property", unlike, say, wealth. In this context, Ian Carter's (1999) approach has been particularly influential. He argues that a person's degree of freedom is measurable, and that it is determined by the number and *extensiveness over space-time* of the actions available to them.

I wish in this chapter to argue that Carter's approach is methodologically and substantively inferior to the axiomatic approach followed in social choice theory. I should clarify at the outset that Carter's approach is only one of two dominant approaches to the measurement of freedom in philosophy; some philosophers (such as Taylor, 1985; G. A. Cohen, 2011; Sen, 1990c; Norman, 1978; Arneson, 1985) see the project of measuring freedom as one of measuring the *value* of people's opportunities, rather than as one of measuring how *extensive* these opportunities are. My critique is not addressed to this second group. The reason for my focus on Carter's work is that, like Carter, I think it is the extent of people's opportunities that is at issue in the measurement question, it is his sort of approach which I believe deserves scrutiny.

I shall proceed as follows. In section 1, I present and motivate Carter's view about the nature of freedom and its measurement. My exegesis will proceed in two steps: first I will explain why Carter thinks we need a measure of freedom and what having one is suppose to solve for us. Second, I will present the actual substance of Carter's proposed measure of freedom. I then contrast this approach in section 2 to the axiomatic approach taken in welfare economics. Here I will highlight three significant methodological advantages of the axiomatic approach: it is more general, it forces us to state our assumptions precisely, and it is more fruitful. In section 3, I take issue with the substance of Carter's view, arguing that the central role that the metaphysics of action plays in his proposal leaves us unable to appreciate the link between the diversity of one's options and one's freedom. Here again the contrast with welfare economics is instructive, and speaks in favour of the axiomatic approach. I devote section 4 to responding to Carter's objections to the approach I favour.

1.1 Carter's measure of freedom

1.1.1 Aims

Most discussions of freedom of choice or social freedom take MacCallum's (1967) analysis of the concept of freedom as their starting point; so does Carter (1999, p. 15). According to MacCallum, when we say that someone is free in some respect, we have in mind the following triadic relation: x is free from y to do/be/have/become z, or x is free, with respect to y, to do/be/have/become z. The z parameter denotes agent x's *opportunity set*, that is, the set of all mutually exclusive options that are not ruled out to x in the present context by the various preventing conditions denoted by the y parameter.

Different conceptions of freedom may be supplied by different interpretations of the parameters. Whether we interpret x as denoting an individual or a group agent will give us a conception of individual or group freedom, for instance. And different conceptions of individual freedom turn primarily on which constraints are believed to be relevant. The primary difference between what List and Valentini (2016) refer to as "freedom as independence" and the so-called liberal model of freedom defended by Carter, Oppenheim (2004) Steiner (1983), and others is that liberal theorists consider that one is free to do x just in case no one *actually* interferes with one's x-ing, whereas independence theorists consider that one is free to do x just in case there is a *robust* absence of externally imposed constraints on one's x-ing, meaning that not only are there no constraints on our action, but no one can easily impose such constraints (i.e., the nearest possible world in which someone does impose constraints is quite distant). Capability theorists like Sen (1990c), meanwhile, consider that one is free to x just in case one can, in fact, x if one wishes to – i.e., just in case there are *no* constraints *at all* that prevent one from x-ing.

Among theorists of the same school there are disagreements over which constraints are relevant: Oppenheim (2004, p. 278) argues that punishment and physical prevention are two distinct potential causes of unfreedom. Moreover, he argues that the unfreedom to do a specific thing is a matter of degree – an unscalable wall restricts my freedom of movement to a greater degree than a merely tall wall that an exceptionally fit individual might climb. In contrast, Carter (1999, p. 219) insists that physical prevention is the only (necessary and sufficient) possible cause of one's being unfree to do something, and, moreover, that this is not a matter of degree: with respect to any particular action x, one is either free or unfree to do it.

With this in mind, supposing we've agreed on a conception of individual freedom, what would it mean to measure an individual's overall degree of freedom? And why should we want a measure of freedom? A number of writers object that the project of "measuring" a person's overall degree of freedom is incoherent. Oppenheim (1981, p. 172) asserts flatly that "there is no such thing as [overall] freedom," by which he means that there is only the freedom to do specific things, with respect to particular agents who do not prevent us from doing them, but can be no meaning to the claim that one enjoys more-or-less freedom, *simplicter*, than another. Oppenheim defends this view on many grounds; firstly, he notes (Oppenheim, 1981, p. 418) that different people can be more-or-less free to do specific things than others, and he is simply skeptical that trade-offs of specific freedoms against other specific freedoms can be assessed in such a way that we can meaningfully claim, say, that Canadians are freer than Americans. To put things more precisely, Oppenheim seems to take the view that a certain kind of *unit comparability* across freedom-types is needed to make sense of the notion of overall freedom.

Secondly, Oppenheim's skepticism seems partly grounded in an ontological thesis: "social unfreedom (and freedom as well) is not a "thing," nor "being unfree" a property." Carter (1999, pp. 24-25) interprets Oppenheim here as saying that freedom cannot be measured because it is not a particular – something located in space-time with measurable properties, e.g., a hubcap – nor a measurable property of particulars, like mass or caloric content. Kristjánsson (1996, p. 11) agrees with Oppenheim on this point, arguing that "being free" isn't a property because it is really shorthand for some comparative relation, in just the same way that "being tall" is not a property, but rather shorthand for being taller than so-and-so. Likewise, the implicit metaphysical assumption in Dworkin's putative *reductio* that freedom cannot be measured because it is not a commodity seems to be that only particulars come in measurable amounts, and freedom is no particular. Or perhaps another way to put Dworkin's point is that freedom must come in units of some kind, like commodities, if one is to make sense of the notion of having more or less freedom.

Essentially, the challenge these objections present is the following: if you want to assert things of the sort "Canadians are freer than Americans," you must point to some *extensive quality* of Americans and Canadians which the Americans have more of, their having more of which explains and justifies your assertion. Carter answers this challenge by arguing that freedom really is an extensive quality – meaning that, as Carter (1999, p. 184) puts it: "it has an empirical counting procedure. It is on the basis of the counting of individual units which have been demonstrated to be of equal size (...) that we can say that one object possesses more of', or 'much more of', a certain extensive quality than another object." Thus, in just the same way that one's tallness consists in one's having more-or-less of an extensive quality (vertical extension), Carter (1999, p. 26) argues that one's degree of freedom really does consist in one's having more-or-less of some extensive quality, and therefore that comparing which of two individuals has more freedom comes down to measuring how much of this attribute each has. Again, Carter seems to buy into this idea that freedom must come in units if it is to be measurable. His substantive view, roughly, is that one's degree of freedom is determined by how many specific actions one can perform and how extensive these are over space-time.

Why should we be interested in measuring anyone's overall degree of freedom? A number of writers (e.g., R. Dworkin, 1977, and Kymlicka, 1990, pp. 145-151) have argued that freedom has no value as such, and so we have no interest in measuring how much of it anyone has; rather, we only care about whether individuals enjoy specific valuable freedoms, such as freedom of speech, conscience, assembly, etc., which we care about protecting because those are the freedoms that enable people to live a good life, or because it is only by protecting those freedoms that the state can show equal concern and respect for all citizens. This is what Carter calls the "specific-freedom thesis," and he counters it by arguing that freedom as such *is* valuable; more precisely, he argues that freedom is "non-specifically" valuable – meaning that we have reasons for valuing freedom that do not reduce to reasons for valuing the freedom to do any specific thing – and that this immediately motivates our interest in measuring it (Carter, 1999, p. 31). Note that to say that freedom is non-specifically valuable need not imply that it is intrinsically valuable: money is surely only instrumentally valuable, yet we can value having more money without having yet any clear idea of what use we shall make of it, thus we may value money non-specifically.

Carter offers the following four arguments in support of his claim that freedom is non-specifically valuable. First, picking up on an example due to Berlin (1969), he observes that the joy a prisoner feels upon being suddenly released, or that a people feel upon overthrowing their oppressors, is well explained by imputing to them a desire for freedom and the belief that they are now free to do many things, but it is *not* necessary to impute to them the belief that now there is something *in particular* that they value and are now free to do. As Carter (1999, p. 32) puts it: "they value 'being free to do things' in a general, rather than in any specific sense."

Second, he points out (50-54) that freedom is non-specifically valuable in part because of its instrumental value: freedom is an important means for realizing valuable ends, but because our ends change, and we may be uncertain about what our ends will be, we may care about having a substantial degree of freedom: when the time comes, we will be free to pursue our ends, whatever these turn out to be. Similarly, freedom as such may be a necessary means for achieving important social goals – here Carter (46-50) has in mind most particularly Mill's famous arguments that freedom of speech is the best means of arriving at and disseminating the truth, and that the strongest protections on the individual pursuit of liberty are the surest means of promoting the general welfare.

Third, Carter argues that, although liberalism is committed to affirming the special value of certain specific freedoms, core liberal commitments also seem to imply a concern for freedom as such. Fixing for a moment longer on the instrumental value of freedom, Carter (54) notes that while some liberal political theorists such as Kymlicka claim that only the freedom to perform specific actions that are themselves meaningful for individuals can be valuable, we are often not in a position to say which options are meaningful to whom. Indeed, the central liberal commitment to neutrality between conceptions of the good rather seems to prohibit us from ruling on which freedoms are valuable. But then it seems liberalism is required to see freedom as being non-specifically valuable.

Fourth, Carter argues that the value liberal theorists place on autonomy commits them to the view that freedom is non-specifically valuable. Carter approvingly cites Crocker (1980, p. 115), who argues that freedom is a "non-causal necessary condition" for what he calls "the autonomy complex," which is the valuable capacity to engage in certain behaviour "in many cases where their consequences are negligible or even unfortunate", including "many acts of risk taking, holding to principle, sacrificing, compromising, admitting mistakes, and struggling through." And as Carter (1999, p. 59) notes the behaviours Crocker lists can "cover any aim or action," so to the extent that freedom *is* necessary for the capacity to engage in such behaviours, it is freedom *as such* that is required, not merely freedom to do some specific things. Thus, liberalism requires us to value freedom non-specifically, because freedom is a necessary component of a good that liberalism regards as intrinsically valuable.

Note that Carter's view here is echoed by Raz (1988, pp. 408-409) when he argues that it is necessary to have at least an adequate range of options to choose from in order to live a good life, for without such freedom it is impossible to live an autonomous life, a life of one's own making. Hurka (2011, p. 143) makes similar points, arguing that our way of life may be important to us because it is the way of life we *chose* – in exclusion of others; it is valuable to have choices so that we may say "yes" to this, but "no" to that, etc.. Indeed, the view goes at least as far back as Mill (1885, chapter III, paragraph 3), who writes: "the human faculties of perception, judgement, discriminative feeling, mental activity, and even moral preference, are exercised only in making a choice." For Mill, freedom is valuable because it is necessary for the development of the valuable human faculties that make life worth living. Carter even suggests that the early Rawls (1971, p. 440) held this view because Rawls took liberty to be one of social bases of self-respect –

one of the primary goods in his theory of justice. So Carter is in good company, and this serves to motivate his project well enough: since freedom is non-specifically valuable, it would be nice if we could measure how much of it people actually have, especially since we often talk as though people need *some degree* of freedom.

1.1.2 Substance

Carter (1999, p. 170) defends what he calls the "empirical approach" to measuring freedom, according to which "the extent of my freedom is a function of the extent of action available to me." He opposes this to the value-based approach, according to which the extent of a person's freedom at least partly depends on the value of the actions available to them. As the starting point for his discussion, he takes Hillel Steiner's (1983, p. 74) proposal that agent i's degree of freedom is "equal" to:

$$\frac{F_i}{F_i + U_i} \,,$$

where F_i is the number of actions that i is unprevented from performing and U_i is the number of actions that i is prevented from performing. The intuition behind this formula is that you are freer the more actions you can perform, and less free the more actions you are prevented from performing, with the understanding that $F_i + U_i$ is not necessarily the same for everyone. Of course, this assumption on its own doesn't suffice to justify the precise functional form of Steiner's formula, but let us bracket that issue for now – we will return to it in the next section – and attend first to Carter's response.

Carter, curiously, takes no issue with the substance of Steiner's measure – in particular, he doesn't question whether Steiner's formula is the only or uniquely attractive "empirical" measure of freedom. However, he argues that the metaphysics standing behind Steiner's analysis are in need of clarification and elaboration. In particular, Carter takes the central issue for Steiner's approach to be the individuation of acts, because depending on how acts are individuated it might turn out that any individual can perform an infinity of acts, which would make Steiner's formula unworkable. And because Carter takes Steiner's ranking rule to be substantially correct, it follows that unless we have some canonical method of individuating acts in a way that avoids the problem of infinite acts, individual freedom is not measurable.

The problem of infinite acts is a recurring problem in the philosophical literature on freedom of choice (see e.g., Steiner, 1983 and Kramer, 2003, chapter 2, for their discussion of the problem). In Carter's presentation, the problem arises out of a certain view in the metaphysics of action defended by Goldman (1970), according to which acts are individuated by the act-properties they instantiate (e.g., the property of moving one's arm); thus if two different act-properties are instantiated, two different acts have been performed. But notice that by moving my arm, I can instantiate multiple act properties (e.g., that of extending my arm out of my window, or that of signalling for a turn); on Goldman's view, this means I have performed multiple acts at once. And here Carter's worry is simply that anything that we do may fall under an indefinite number of descriptions, thus instantiating an indefinite number of act-properties, perhaps an infinity, and so it is that Steiner's formula would become useless, since the values of the numerator and denominator would become arbitrary.

In order to counter this conclusion, Carter appeals to Davidson's theory of action, which is the principal rival of Goldman's theory. On Davidson's view, every act that we ever perform is in fact identical with some bodily movement of ours. That is, every act that we ever perform, says Davidson, is either the act of moving our body in some way, or an act that we perform *by* moving our bodies in some way, and because Davidson holds that any act that is performed by performing some other act is identical to the act by which it is performed, it follows that the only acts are bodily movements (Davidson, 1980). Thus, on Davidson's view, if by extending my arm, I extend it out of a window, and signal for a turn, I have not done *three* things, but just one thing, of which three descriptions have been given. This solves the problem of indefinitely many descriptions; counting how many acts an individual is prevented or unprevented from performing comes down to counting how many bodily movements they are prevented or unprevented from making (Carter, 1999, p. 176).

This method of act-individuation nonetheless raises a new problem for Carter. As he writes (178): "[i]f my available bodily movements have counterfactual consequences which yours are prevented from having, we should like to say that I have more freedom than you in this respect." But if actions are identified with bodily movements, the consequences of our bodily movements cannot count as additional actions available to us. To get around this problem, Carter proposes that either we simply count all the foreseeable consequences of our bodily movements *as if* they were additional acts that we perform by moving our bodies (179), or, what is equivalent (according to Carter), in counting how many bodily movements are available to a person, we simply give acts that have longer foreseeable chains of consequences greater weight in Steiner's formula, in proportion to how extensive over space-time are the events of which the act is the bringing-about (80). Thus, for example, if I can cause an avalanche by clapping my hands, whereas you cannot, then the act of clapping my hands grants me more freedom than the act of clapping your hands grants you in proportion to how much larger a region of space-time is covered by my hands as they clap and the avalanche as it rolls down the mountainside than is covered by your hands clapping (205). Since Carter believes these two solutions to be equivalent (for reasons that need not detain us here), I shall focus on the second solution as a description of his view.

One qualification here is needed. It is occasionally the case that two different possible acts would bring about very similar events, which would have a degree of overlap in their space-time coordinates. For instance, firing a gun with my right hand and firing a gun with my left hand may both result in a bullet's tracing a particular trajectory through space-time and killing a man (193). Here Carter argues that it would be to double count if we were to weigh each of these two different acts by the physical extension of the events which they foreseeably cause (189). Intuitively, firing a gun with my left hand and firing a gun with my right hand seem like two ways of doing the same thing – two ways of bringing about the exact same sequence of events. Thus, in order to avoid double-counting, different weights must be assigned to acts which overlap in this way (190); in particular, if any number of acts overlap, only one of these gets weighted by the extensiveness of the overlap region.

Additionally, to avoid running back into the problem of infinite acts once it is noted that the infinite divisibility of space-time implies an infinity of possible bodily movements, Carter (184) suggests that we should measure space-time *as if* it were granular. More precisely, he suggests that we should pick a smallest unit of time and a smallest volume unit of space, thus obtaining a coarse-grained grid of space-time; acts are then individuated by the regions of this grid which they occupy, leaving each individual with only a finite number of acts to perform.[1] In sum, according to Carter, the degree of freedom I enjoy is given by the ratio of the number of "paths" (i.e., acts) I am unprevented from tracing across this coarse-grained space-time grid by moving my body to the total number of paths I could trace were I not prevented from doing anything, where each path is weighed (avoiding double-counting in cases of overlap between different acts) in proportion to the physical extension of all the events foreseeably caused by tracing this path with my body.

[1] It's not clear why Carter had to resort to this coarse-graining idea. If the weight that any act is to have in the measure of overall freedom is anyway identical to its physical extension plus the physical extension of all the events that it foreseeably causes, why couldn't we dispense with the idea of counting acts altogether, and simply declare that one individual's degree of freedom just is the area of space time that they and the events they could cause could ever occupy? We don't need space to be granular, after all, to measure its volume; nor time, to measure its length. Double-counting wouldn't be an issue here either.

Naturally, Carter doesn't expect that we go out in the world to make the physical measurements needed to determine the value of each individual's Steinerian ratio; instead, he suggests that good proxy measures for an individual's extent of overall freedom exist. For instance, he claims that the exchange value of an individual's resources holdings provides, plus the market value of their human capital and other marketable skills, provides us with a good indicator of their freedom, because more money generally translates to more available acts (1999, Chapter 10, section 3). This measure is relatively easy to apply, but is obviously something of a dummy measure, so Carter suggests that it can be refined by being combined with an adapted version of the UNDP's freedom-type metric.

First, Carter instructs us to develop a list of different specific types of freedom – a list which he insists must be morally neutral, in the sense that the list must be drawn up by some non-moralized selection procedure, and not limit itself merely to those freedom types we happen anyway to think are especially valuable – and then to weigh each freedom type by the share of one's empirical freedom that corresponds to the exercise of this freedom. More precisely, real numbers are to be assigned to each freedom type, indexed to each individual, which are strictly increasing in how large a share of one's empirical freedom falls under the exercise of this freedom-type, and strictly decreasing in the extent to which the exercise of this freedom is limited, and the likelihood of this freedom's exercise being limited (Chapter 10, section 2). These numbers are to sum to 1 in the limiting case where there are no preventing conditions on the individual's freedom, and to 0 when the individual has no freedom along any of those dimensions (286). Carter then suggests that we simply add up the values of each freedom-type, and then multiply the total "score" by the exchange value of the individual's resource holdings (286). Carter (286) explains this last move as follows: "[th]e basic intuition here, is that we should first look at the agent's personal resources – itself an indicator of various preventing conditions or their absence – and then go on to ask what the agent is prevented and unprevented from doing *with those resources*." We will have a critical look at the substance of this proxy measure later on, but for now, let us simply note Carter's claim that it will approximate the individual's overall freedom.

Carter believes the Steinerian analysis of overall freedom has a number of attractions. For one thing, he argues that Steiner's formula (suitably interpreted in light of the above considerations) turns out to generate the comparative assessments of overall freedom we find most intuitive. For instance, he argues that this analysis will turn out to imply that having more qualitatively varied acts open to us makes us freer, because the events that qualitatively similar acts bring about will generally overlap spatiotemporally to a much greater degree than the events which dissimilar acts bring about. And thus, in light of the no double-counting requirement, acts will be given greater weight in the Steinerian formula if they are dissimilar from existing alternatives. Furthermore, he contends that this formula has the great advantage of making freedom cardinally measurable (indeed, measurable on an absolute scale), which will in turn permit us to assess which social states maximize aggregate freedom and to measure how equally the aggregate quantity of freedom in the world is distributed. Finally, he argues (for reasons that will be made clear in section 5) that the Steinerian analysis offers the only value-neutral way of measuring individual freedom; that is, it is the only analysis that is consistent with the claim that freedom is non-specifically valuable, and doesn't collapse the measure of freedom into the measure of its conduciveness to the specific values that matter to us independently.

I am skeptical of Carter's conclusions, but before criticising them, I turn to a presentation of how the question of the measurement of freedom is treated in welfare economics. I will highlight the three features of this approach which make it, in my view, a preferable approach to the question of measurement than Carter's and Steiner's.

1.2 The axiomatic approach

1.2.1 Method

Welfare economists start out with the same aims as before: we want to way to make meaningful claims of the sort: "Canadians are freer than Americans," "Cassandra enjoys a greater degree of freedom than Aurelius", or even "there is a more equal distribution of freedom in Denmark than in Venezuela." But unlike the authors mentioned in section 1, welfare economists do not take the task of making sense of such claims to be the task of finding some natural quantity in the world, one's having more of which somehow makes one more free, and which we have been referring to all along in speaking about freedom; rather, they simply set out to investigate what sorts of principles for ranking sets of alternatives make best sense of our intuitive comparative judgments or are most normatively attractive.

On this approach, we start by defining X, the set of all possible options. X is assumed to be finite, but otherwise no interpretation is imposed on what, exactly, an option is. As Pattanaik and Xu (2015) explain, the way options are most usefully specified may depend on the context: we may choose to think of an option as a bundle of actions that an agent may perform, or as a pair composed of a bundle of actions and their attending consequences, or we may choose to specify options as propositions that individuals can make true, and, depending on our interests, we may also be satisfied with specifying options at a greater or lesser level of granularity. For instance, we need not always care about how many verbal utterances individual can make, but only whether they can criticize the government.[2]

Now we define $Z = 2^X \setminus \varnothing$ as the set of all opportunity sets, where an opportunity set is understood as the set of all mutually exclusive options that are available to the individual (i.e., not ruled out to them by whatever set of constraints is considered relevant). Of course, one can consider whatever constraints one wishes to be relevant in determining the content of each individual's opportunity sets, so any substantive conception of freedom (independence-based, liberal, capability-theoretical, etc.) can be applied to fill in the details. It is assumed that an individual's degree of freedom is determined by their set of available alternatives, we define $\succsim$, a transitive and reflexive binary relation on Z, with $\succ$ and $\sim$ denoting the asymmetric and symmetric parts of the relation, respectively. We interpret $A \succsim B$ as "A offers at least as much freedom as B." The aim now is to propose intuitively plausible axiomatic constraints on $\succsim$ that will deliver a (preferably complete) ordering over Z. Standardly, a representation theorem is used to show under what necessary and sufficient conditions a ranking over Z satisfying certain axioms implies that one opportunity set offers more freedom than another.

Already we can note important contrasts between the two approaches. In assuming X to be finite, the axiomatic approach assumes away the core worry that motivates Carter's approach. In refusing to impose a particular interpretation of options, it reveals some disinterest in the issue of how, precisely, acts are to be individuated – the issue at the heart of Carter's proposal. Indeed, welfare economists here are in the business of trying to rank sets of options, *however exactly options are conceived of.*

Also, notice that while Carter and Steiner speak of measuring the freedom of individuals, welfare economists speak of ranking sets of alternatives, and this reflects a disagreement about what it means to measure freedom. As Carter (1999, p. 184) himself notes, he takes the measurement of freedom to be the empirical counting procedure of some extensive quality, whereas welfare economists take it to be rather like the measurement of utility, that is, as the attempt to establish a

[2] Of course, if we specify options in this way, at the level of granularity we think relevant, in terms of those features we think are salient, then the specification of the options is clearly value-laden. But we shouldn't make too much of this. For one thing, although we can specify option in a value-laden way, we don't have to. We could also individuate options as Carter wants us to. The axiomatic approach itself imposes no interpretation at all on what an option is, and is in this sense value-neutral.

numerically representable ordering over fairly generally specified classes of objects. This way of framing the project also sweeps away the sorts of concerns voiced by Oppenheim and others: whether freedom is a property or a relation or a particular is irrelevant to the question of what properties the ordering on Z should satisfy. Finally, the locus of disagreement in both approaches is quite different: whereas in section 1 the disagreement was metaphysical (is freedom a property or a relation? Are acts particulars or instantiations of act-properties?), here disagreement focuses on what makes someone more-or-less free.

A few examples will suffice to show this. One debate in the literature concerns the role of preferences in establishing a ranking over Z. Pattanaik and Xu (1998), for instance, propose that for any opportunity sets A, B and some option $x \in X \setminus A$, if some rational person could prefer x to every alternative in A and some rational person could prefer at least one alternative in A to x, then $A \succsim B$ implies $A \cup \{x\} \succ B$. Otherwise, if no rational person could prefer x to any option in A then $A \cup \{x\} \sim A$. This is an example of an axiomatic requirement that options be minimally desirable for their addition to one's available set of alternatives to increase one's freedom. Barberá and Godal (2010, p. 273) suggest yet a very different role for preferences, in the context where it is uncertain which options in the various sets one is ranking will remain available at the moment of choice. On their proposal, one set offers more freedom of choice than another if it has a higher "expected opportunity," meaning roughly that it is the set with the best expected set of alternatives remaining at the time of choice. This is intended to capture the idea that it is best to have flexible sets of options, ones whose ability to provide us with meaningful choice is relatively invariant under revisions in our beliefs about the state of the world.

Other authors still have proposed constraints on $\succsim$ that give the *diversity* of one's options a role in determining how much freedom of choice these options offer. Pattanaik and Xu (2000) and Bervoets and Gravel (2007), and Klemish-Ahlert (1993) all propose different axioms for ranking sets purely in terms of the diversity of their elements, all based on different interpretations of diversity. This line of inquiry has sometimes yielded surprising results: van Hees (2004), for instance, proves a number of impossibility results for ranking rules that attempt to consider diversity. For instance, he shows that if X contains at least 4 elements, no transitive ranking of Z exists satisfying the condition that $A \cup \{x\}$ offers at least as much freedom as $A \cup \{y\}$ if and only if x is more dissimilar from the element of A it is least dissimilar from than y is dissimilar from the element of A it is least dissimilar from. I only mention this result for how surprising it is: intuitively, x adds greater variety to your menu of choices than y, given that it is less similar than y to any of your existing alternatives. And this well illustrates the value of the axiomatic approach: it can prove commonsense intuitions to be untenable, forcing us to refine our pre-theoretical intuitions about what makes a person more-or-less free.

I stated earlier that the axiomatic approach assumes that an individual's degree of freedom is determined by their opportunity set. It has therefore seemed to some (e.g. Carter, 2004) that it cannot take account of the options one is prevented from choosing in establishing rankings over Z. Carter (2004) sees this as a weakness in the social choice literature, a failure to distinguish *freedom*, which makes essential reference to what you are prevented from doing, and *freedom of choice*, which only looks at what you can in fact do. To see that, contrary to Carter, it can, suppose that we define for every set $A \in Z$ an "anti-set" $A^* \subseteq X \setminus A$, which, intuitively, corresponds to the set of all acts one is prevented from performing. A^* need not be assumed to be A's complement over X, as there may be acts that one is unable to perform even though one has not been prevented from doing them. We can then define:

Monotonicity-1: $\forall A, B \in Z$, $|A^*| = |B^*|$ and $|A| > |B|$ imply $A \succsim B$

Monotonicity-2: $\forall A, B \in Z$, $|A| = |B|$ and $|A^*| > |B^*|$ imply $B \succsim A$

1,2-Invariance: $\forall A, B \in Z$, $B \sim A \iff |A| = |B|$ and $|A^*| = |B^*|$

In brief, we can require individual freedom to be increasing in the size (or diversity, etc.) of one's opportunity set, and decreasing in the size of one's anti-set. It is of course striking that axioms like Monotonicity-2 and 1,2-Invariance are not really discussed in welfare economics. This may be because axioms like Monotonicity-2 are inconsistent with an axiom that has found near-universal acceptance in the social choice literature, namely what Pattanaik and Xu (1990) call "indifference between no-choice situations," which requires that for any two x, $y \in X$, $\{x\} \sim \{y\}$. The intuition behind this axiom is simply that if you have no choice about what to do, if you're limited to doing just the one thing you can do, then you just don't have any freedom at all. Freedom requires choice. But as it is not necessarily the case that $|\{x\}^*| = |\{y\}^*|$, this axiom rules out any ranking of opportunity sets that depends on the size of these sets' anti-sets. Of course, this can be read as a *reductio* against the indifference axiom,[3] but it may explain why anti-sets aren't usually considered.

[3] I see the force of the *reductio* as running in the opposite direction. Indifference between no-choice alternatives is beyond compelling: it commands a Moorean certainty. Anyone who would deny that two individuals each lacking any possibility for choice are equally (un)free has, for my money, *ipso facto* blown a hole in their theory of freedom.

1.2.2 Initial attractions

More could be said about the substance of particular ranking rules, but for our purposes it will do well enough to have noted the core features and assumptions of the axiomatic approach. I now outline three reasons why the axiomatic approach is a more promising way of investigating what it is to be more-or-less free and who has more freedom than the sort of approach taken by the theorists mentioned in section 1.

Generality

First, because it imposes no interpretation on the concept of an option, it is more flexible in its possible range of applications than the empirical approach. If we are interested in comparing individuals' degrees of overall freedom, we can think of options as mutually exclusive lists of acts, where each item in the list is a different sort of act. We can make the list longer or shorter according to how much specificity we wish to pack into our comparisons. Then we run our preferred ranking rule, and we obtain a ranking of opportunity sets that will tell us who is freer than whom.

But suppose we found the specific-freedom thesis compelling, and that we were only interested in comparing the extent of individuals' freedom of speech, or freedom of worship, or other valuable sorts of specific freedoms. On the axiomatic approach, there are fairly natural ways of making such comparisons: in the case of freedom of speech, we could start by partitioning the range of possible speech acts more-or-less coarsely into different types of speech – e.g., anti-government speech, academic speech, religious speech, small talk, speech calling for collective action, conspiracy theory speech, pornographic speech, hate speech; printed speech, public speech, visual speech, etc. – and then conceive of options as speech act types or bundles of speech act types. Each individual's opportunity set is then determined by asking what speech act types they can perform, and then we can apply whatever ranking rule we feel is best to compare who has more freedom of speech than whom.

The empirical approach, by contrast, demands that we find some extensive quality that can be measured by an empirical counting procedure, the having more-or-less of which determines who has more or less freedom. It is not obvious how to extend this analysis to specific types of freedom. Take again the case of freedom of speech. Speech act types are not extensive qualities; they don't come in equally sized units of stuff measurable by physical instruments, so something else would be required. And it would be a hopeless to list every speech act *token* than every individual could ever physically perform. This list may well be infinite, and is certainly not enumerable in a human lifetime. Carter's proxy measures have no clear application either. Which fraction of the exchange value of an individual's total resource holdings should we take to be indicative of the extent of their freedom of speech? If exchange-value metrics tell us anything at all about how free an individual is, they can only point to how free they are overall, not how extensive any of their specific freedoms are.

Carter's second proxy measure doesn't seem to me to be more helpful. Recall that here the idea is to assign numbers to each freedom type according to how large a share of one's empirical freedom corresponds to the exercise of this freedom and to how far the exercise of this freedom is restricted, then to multiply their sum by the individual's exchange-value score. This is all a bit perplexing. Carter tells us that multiplying the freedom-type score by the exchange-value score is supposed to tell us how much you can do with your resources. But surely the freedom to do things with resources is itself a kind of freedom, or is included in other types of freedom (whether you can or cannot afford a car may be relevant to our assessment of your freedom of movement), so the extent of what you can do with your resources must already have been assessed in the computation of the initial freedom-type score. Perhaps Carter would disagree, but it's hard to say since he doesn't provide us with a list of freedom-types, merely insisting that the list should be drawn up in a value-neutral fashion. Moreover, even if we can avoid the appearance of vicious double-counting, why are we multiplying? It's certainly not self-evident that *this* is the operation which tells us how much you can do with your resources. Why not add the two scores, or take the factorial of their sum, or exponentiate one score by the cubic root of the other?

But let's bracket concerns over multiplication, and focus only on the freedom-type metric. Carter doesn't say much about how one should estimate the share of overall empirical freedom different freedom types take up, so how does one apply this method? As far as I can see, just in the case of free speech, we would be required to guesstimate, for every speech act we could ever perform (possibly an infinite list!), what bodily movements of ours would be identical with said speech act, and what the counterfactual consequences of all those speech acts would be, then to measure the extensiveness over space-time of these movements and consequences, and finally to determine the Steinerian ratio of unprevented acts to unprevented + prevented speech acts. This is a heroic task.

Of course, everything I've said here about freedom of speech goes for other freedoms as well. The axiomatic approach offers us natural ways of establishing comparisons in individuals' degrees of freedom of worship, association, assembly, etc. But again it's hard to see how the empirical approach could be applied to establish comparisons of the extent of specific freedoms; it seems limited to making comparisons of overall freedom. And yet if there are especially valuable types of freedoms, we might care about being able to make comparisons in degrees of specific freedoms. The axiomatic approach, then, promises to serve more of our theoretical aims than the empirical approach, and this speaks in its favour.

Precision

A further methodological advantage of the axiomatic approach is that it allows us (forces us, even) to state our assumptions precisely. The benefits of this have already come through in the discussion above. For one thing, as Van Hees has shown, attempts to articulate precise criteria for what makes one set of options more diverse than another may force us to reconsider our intuitions about diversity. This point holds quite generally: because the axiomatic approach works by proving representation theorems, which establish that a binary relation on Z satisfies certain axioms if and only if it ranks Z according to some particular rule, the logical implications of our assumptions can be precisely studied, which makes it easier to know which assumptions to weaken or reject. For instance, it is because Pattanaik and Xu's (1990) original (and apparently plausible) monotonicity and composition conditions implied (absurdly, in their view) that A offers more freedom than B just in case $|A| > |B|$ that they later come to reject these conditions. And likewise the next chapter argues for the rejection of an assumption that most theorists working within this approach have adopted.

More importantly, however, in *not* stating our assumptions precisely, we run the risk of making conceptual errors, such as the error of confusing a scale of measure that cardinally represents freedom (i.e., it allows comparisons in degrees and differences in degrees of freedom) with a scale of measure that is itself cardinal but represents freedom merely ordinally (i.e., it measures *something* cardinally, but only allows comparisons in degrees of freedom, not degrees and differences in degrees).[4] And here I wish to scrutinize Carter's claim that Steiner's formula, appropriately interpreted, offers a cardinal representation of overall freedom.

[4] This sort of error is familiar to us from the literature on the measurement of preference. Suppose we assume that what is good for a person is to have their preferences satisfied, so that the more their preferences are satisfied, the better-off they are. Von Neumann and Morgenstern proved that if individual preferences satisfy certain conditions, then a cardinal representation of those preferences exist. You might think this shows that personal goodness is therefore also cardinally measurable. But as has been widely noted (see Ellsberg, 1954; Arrow, 1963; Sen, 1976; J. A. Weymark, 1991; Broome, 1991b), the mere fact that a cardinal representation of individual preferences exists does not mean that this is the *only* representation of their preferences that exists. The scale (utility, in the case of preferences) is cardinal, but it need not represent cardinally what it represents.

As mentionned previously, Steiner claims that each individual's degree of freedom is "equal" to the value of their Steinerian ratio. In fact, however, it seems to me that his formula only offers us an ordinal ranking rule. That is, his formula can be used to order individuals by how free they are, but differences in Steinerian scores have no meaningful interpretation. Recall that Steiner justifies his choice of formula by invoking the principle that one can only be more (less) free to the extent that (1) one can perform more (fewer) acts and (2) is prevented from performing fewer (more) actions. Steiner presents his ranking rule as though it is the only rule which satisfies (1) and (2), but in fact there are in fact many different ways of ranking individuals in terms of how much freedom they have that are consistent with assumptions (1) and (2).

For a start, any positive monotone transformation of Steiner's formula will rank individuals by their degrees of freedom in exactly the same way, and so trivially satisfy (1) and (2). Moreover, there are many other rules which induce different orderings but still satisfy conditions (1) and (2). For instance, we could just as well say that i is freer than j just in case $F_i - U_i > F_j - U_j$. This second ranking rule is consistent with (1) and (2), but whereas Steiner's rule implies that there is an upper limit to how free ($U_i = 0$) or unfree ($F_i = 0$) one can be, the rule I've just proposed does not. Or else, consider the rule according to which i is freer than j if and only if:

$$\frac{F_i}{1+U_i} > \frac{F_j}{1+U_j}$$

In this case, there is a lower bound on how free any person can be ($F_i = 0$) but no upper bound. This might sound plausible. Matthew Kramer (2003, p. 359) proposes yet another formula that is identical to Steiner's, but where the numerator is squared, arguing that since we should care more about what people can do than what they are prevented from doing, it is necessary to square the numerator to make increases in in people's exerciseable liberties sufficiently more weighty in the calculation of how free they are than increases in the number of things they are prevented from doing. Kramer's precise choice of exponent is arbitrary, but his formula is consistent with the restrictions laid down by Steiner. In fact, the only ranking rules that *are* ruled out by (1) and (2) are ones which permit for an agent to be more free than another even though they are unprevented from performing no more actions and are prevented from performing no fewer. In other words, (1) and (2) on their own only imply a dominance-based partial order, which is not terribly restrictive.

Steiner himself does not argue for an upper or a lower bound. He takes it for granted that having more available actions must expand one's degree if freedom, but does not claim that there is an upper bound to this. Likewise, he insists that one should take account of the number of acts one is unfree to perform in assessing an individual's degree of freedom, and not "confine our calculation to summing those actions which [they are] free to do" (Steiner, 1983, p. 74). Otherwise, we would be guilty of confusing freedom with ability, says Steiner, and though an individual may be free and able to perform more acts than another, they may also be able but unfree to perform far more as well. To simply ignore the number of acts the individual is unfree from performing in assessing their degree of liberty "is to misconstrue the object of such an exercise" (75). Kramer (2003, pp. 360-368) argues likewise that there's just something special about humanly imposed constraints that calls out for attention: a person who cannot walk more than 100 meters because others systematically assault her is unfree in a way that a person who cannot walk more than 100 meters due to disability is not, and this form of unfreedom must be represented in our measure. These arguments have some force, but they are plainly just arguments for condition (2). Something more would have to be said to establish that two individuals who are both unprevented from performing any act, though one is free to perform far more, are equally free.

Maybe Steiner's formula offers the best interpretation of the relation between individual freedom and extents of available action, but there are substantive questions to be settled in deciding between different ranking rules. If I am prevented from performing exactly as many acts as you, but am unprevented from performing significantly more, aren't I freer than you? This certainly sounds intuitive. Well, according to the rules I proposed, then I am necessarily freer than you in the case described, but not so on Steiner's (if $U_{me} = U_{you} = 0$). Or else, from the moment that we accept that my degree of freedom decreases as the number of actions I am unfree to perform increases, why must there be an upper bound on how much my freedom can be made to decrease in this way, as Steiner's rule implies?

Carter doesn't consider these issues. He just asserts that Steiner's measure, suitably interpreted, offers a cardinal measure of freedom. But as we've now seen, Steiner's measure only offers an ordinal measure of freedom: any positive monotone transformation of it generates the exact same ordering. For Steiner's formula to represent individual freedom cardinally, additional assumptions are required to those which he imposed. Steiner himself does not make any such assumptions explicit, so perhaps we should turn to Carter for illumination.

Carter talks at various points of freedom "being a function" of something or other, and, in particular, he defends the view that "the extent of my freedom is a function of the extent of action available to me" (Carter, 1999, p. 170). In explaining why freedom, so conceived, is cardinally measurable by Steiner's formula, he says:

> "The idea of measuring 'extents of available action' in terms the physical dimensions of the events agents are free to bring about allows, in theory, for absolute, cardinal measurements. This is because space, time, and matter are, as we saw in Chapter 7, 'extensive' qualities, and therefore ones on which we can in theory perform concatenation operations so as to be in a position to say 'how much' of those qualities is present." (270)

I take it that Carter is arguing the following: Steiner's formula (suitably interpreted) measures one's freedom cardinally because (i) it offers a correct account of the relation between one's degree of freedom and the extent of one's available action, and (ii) one's extent of available action is cardinally measurable.[5] This reading seems to fit with Carter's talk of freedom being "a function of" one's extent of available action. If this is the right reading of the above passage, however, Carter is simply mistaken. If all we assume is that freedom is strictly increasing in the extent of one's available action, then even if the extent of available action is cardinal, Steiner's formula does not measure freedom cardinally.

To see why, consider the following analogy: suppose individual welfare is strictly increasing in their level of income. Income is cardinally measurable, but even if we assume that income is the sole component of individual welfare, it does not follow that to measure people's income is to measure their welfare cardinally. If all we know is that people's welfare is an order-preserving function of their income, then all we learn from the fact that I earn $10 and you earn $20 is that

[5] A different way of reading this passage is that Carter is *identifying* freedom with extents of available action, and telling us that freedom is cardinally measurable because extents of available action are cardinally measurable, in the same way that temperature is cardinally measurable because temperature *just is* mean molecular kinetic energy, and mean molecular kinetic energy is cardinally measurable. In other words, when we speak of a person's freedom, or their degree of freedom, and when we make claims like "Canadians are freer than Americans," we are actually obliquely referring to a ratio of regions of space-time. On this reading, Carter would be (trivially) correct that Steiner's formula measures freedom cardinally.

This is not a very plausible reading; "freedom" and "the ratio of acts one is unprevented from performing to the sum of the acts one is unprevented from performing and the acts one is prevented from performing" are not synonymous expressions, nor is "freedom" a proper name for a physical magnitude (like "temperature"), so why should we believe they co-refer? Ordinarily, when two non-synonymous expressions are discovered to co-refer, it is because these expressions are definite descriptions which pick out the same object on the basis of different properties; "the morning star" and "the evening star" were discovered to co-refer when it was found that the first light to shine in the night sky was also the last light to go out. But "freedom" isn't a definite description of ratios of areas of space-time; it's a somewhat vague concept which intuitively has to do with how capable one is to do, be, have, become things that are worth wanting. But there are different interpretations of the relation between one's freedom and the set of things one can do; this is why I think Carter and Steiner are best interpreted as offering an interpretation of this relation.

you are happier than I – we do *not* learn that you earn twice as much well-being as I do, though you earn twice as much money. Steiner's formula could only be taken to measure freedom cardinally if it was assumed that freedom is not only increasing, but also strictly linear in the areas of space-time one can influence, and why should we believe that? It would contradict commonsense intuitions about the importance for freedom of having a diversity of nontrivial options (Raz, 1988, p. 376). And mightn't freedom, like welfare with respect to money, exhibit diminishing marginal returns to influenceable areas of space-time?

Fruitfulness

A final methodological advantage of the axiomatic approach is that it focuses our attention where it needs to be: on substantial questions about what principles are most normatively appealing as interpretations of the "offers at least as much freedom as" relation, rather than on largely orthogonal questions about ontology or the metaphysics of action. As we saw earlier, Steiner and Carter neglect alternate ways of measuring individual freedom, even though many other ranking rules are consistent with their assumptions, and there are substantial questions to be settled about which rule is best. In contrast, this is where all the critical attention is focused in welfare economics, with the predictable result that a wilderness of different ranking rules have emerged, all expressing different commitments about the role of preferences and diversity in the measurement of freedom.

And this diversity of views is surely a good thing. It is a serious question which sets of axioms are most compelling, and if different sets of axioms are all compelling in different ways, we must decide whether to come down hard on the side of one set of axioms, abandoning some of our intuitions along the way, or to accept that we have many competing but equally adequate conceptions of what makes a person more-or-less free. Even if we think that we ultimately ought to pick one set of axioms, having a wide variety of compelling alternatives to consider may make us less dogmatic about what the correct ranking rule is, it may reveal to us new possibilities we hadn't considered, and the general back-and-forth nature of arguing for one set of axioms against another may help us make better sense of our intuitions and of what matters. Likewise, even if we find one ranking rule on the whole more attractive than others, we may nonetheless have greater confidence in the rankings that it implies when other ranking rules that we find compelling agree with it.

To focus instead on questions in the metaphysics of action is simply not to ask the right questions, *if* what we are interested in is to develop a measure of freedom. Consider some analogies. Would it be appropriate to demand of consumer choice theorists, before they started the business of offering hypotheses, proposing models, making predictions, etc., about individual choices over consumption bundles, that they first settle the question of whether bundles of goods are to understood as substances with properties, or instead as bundles of property intensities and location properties? Would it be appropriate, in formal epistemology, to subordinate the question of how agents should update their beliefs about the possible states of the world to the question of whether these states are to be understood in modal realist fashion, as referring to other actual worlds in parallel universes, or instead merely as propositions expressing counterfactual scenarios, in Kripkean fashion? These may be important questions, but they are separate questions from the ones of interest in each case, and it seems to me that the question that is of interest to us when we are trying to make sense of claims like "Canadians are freer than Americans" is: "what, in general, makes anyone more-or-less free than anyone else?" This is just what the axiomatic approach invites us to do.

1.3 Freedom, Carter, and diversity

In fact, not only are questions in the metaphysics of action not the questions of concern when we are interested in the measurement of freedom, but the role Carter gives to metaphysics in defining his position ends up working as a straightjacket: it forces him to stipulate that an individual's degree of freedom does not depend on the diversity of their available options. This is a problem because, as Carter acknowledges, it seems very obvious that the diversity of one's options has a lot to do how much freedom one enjoys – indeed, he believes the biggest challenge for his view is to recover this intuition. Citing an example from Norman Daniels, he notes that someone with a choice between 21 different brands of detergent does not appear to have significantly more freedom than someone with a choice between 7 different types of detergent. Likewise, citing an example from Pattanaik and Xu (1990, 2000), he notes that it seems false to say that the opportunity set {travel by car, travel by train} offers no more freedom than the opportunity set {travel by blue car, travel by red car}. Yet a naïve reading of Carter's view suggests that this is just what his empirical approach implies, because diversity is not a feature of the spatiotemporal extensiveness of an act.

Carter, however, points out that the 21 brands of detergent only provide us with 21 ways of doing the same thing: the events we can causally generate with one brand of detergent (e.g., washing our clothes) overlap to a very substantial degree with the events we can causally generate with any other brand. Likewise, the set of events that travelling in a blue car allows us to bring about (e.g., moving our body from point *a* to point *b* along route *c*) overlap to a much more significant degree with the set of events that travelling in a red car allows us to bring about than the events that travelling by train do. Thus Carter believes he is able to recover our intuition that one is freer the more diverse one's options, without breaking his "empiricist" commitments.

Unfortunately, I do not think that Carter's analysis takes him very far. Although in the two cases just mentioned, similarity and overlap went together, it is not, in general, either a necessary or a sufficient condition for two acts' being more (less) similar that the events they allow us to bring about overlap to a greater (lesser) degree. One shouldn't therefore expect much covariance between the two, nor in general expect the one to be explained by the other.

1.3.1 Sufficiency

Consider the following example: my experience of many Italian cities has been that the only restaurants within a reasonable price range that one can find are pizzerias and pasterias. Do the options in the set {eat at pizzeria *a*, eat at pizzeria *b*, ... eat at pizzeria *z*} overlap spatiotemporally to any substantial degree? No: the different pizzerias are in different buildings, which occupy different regions of space-time; if I went to eat at pizzeria *a*, I would sit in a different chair, at a different table than I would be at pizzeria *b*, or *c*, use different cutlery, etc.

Compare this choice situation to the choice situation one faces in a highly cosmopolitan city, where multiple culinary traditions co-exist. There the typical person's opportunity set with respect to dining options might be {eat at Korean restaurant, eat at Vietnamese restaurant, eat at Thai restaurant, eat at pizzeria, eat at deli, eat at French restaurant, eat at Irish pub, eat at Korean-Swedish fusion restaurant, eat at Peruvian restaurant, etc.}. Even if there are no more restaurants to choose from in this hypothetical cosmopolitan city than in my hypothetical Italian city, it is perfectly obvious that the menu of choices on offer here exhibits considerably more diversity. Yet the acts of eating at pizzeria *a* and pizzeria *b* do not overlap more than the acts of eating at *Siam Kitchen* and eating at *Seoul Food*; in fact there may be much *less* overlap between the first pair of acts, if the two pizzerias are far apart while *Siam Kitchen* and *Seoul Food* are right next door to each other. Clearly, that two acts overlap to a lesser degree does not imply that they are more dissimilar.

1.3.2 Necessity

Nor is the converse true: dissimilar acts may significantly overlap spatiotemporally. Just consider information technologies. Without turning my nose away from my computer screen, I can watch a movie, read a paper I found on Jstor, grade my students' essays, read about the history of hubcap manufacturing, play chess, listen to music on Youtube, etc. These are all very different activities, requiring different skills, satisfying different interests, inducing very qualitatively different mental states in me, but each activity overlaps significantly over space-time with every other: as far as my bodily movements are concerned, these activities mostly consist of my sitting around in the same spot, and the events that I bring about (e.g., causing some bits to flip from 0 to 1) will tend to overlap in their spatio-temporal dimensions as well.

The extent of spatiotemporal overlap between available acts thus appears to be a poor indicator of their diversity. Given that it is neither necessary nor sufficient for one opportunity set A's being more diverse than set B that there be less overlap in the acts that A makes available to us, one shouldn't expect a very tight connection between diversity and overlap. But if our degree of freedom closely tracks the diversity of our options, then Carter's approach risks seriously misrepresenting who enjoys greater freedom than whom. The axiomatic approach, by contrast, allows us to give the diversity of one's options a much more direct role in our assessments of overall freedom. Many such measures have been proposed (Weitzman, 1992; Pattanaik and Xu, 2000; William Bossert, Pattanaik, and Xu, 1992; Baumgärtner, 2006; Bervoets and Gravel, 2007; Pattanaik and Xu, 2008; Hees, 2004; Nehring and Puppe, 2009), and indeed I will propose my own in chapter 4.

1.4 Carter's objection

If what I have argued so far is correct, then we have reached a provisional case in favour of the axiomatic approach. For this case to be truly probative, however, we must confront the empirical approach on the ground where it is strongest. And indeed, the most important advantage that Carter claims for the empirical approach, in my view, is that it is the *only* approach to the measurement of freedom that is consistent with the thesis that freedom is non-specifically valuable.

Carter attributes the view that a new freedom must be "significantly different" from one's previously available freedoms to contribute significantly to one's overall freedoms to philosophers like Sen (1990), Arneson (1985), and Charles Taylor (1985). Taylor, for instance, argues that freedom is only important to us because we are purposive beings. One person should only count as freer than another if one is free to pursue a greater range of purposes, and more important purposes. Or perhaps more precisely, an additional option only contributes to our freedom if it makes the pursuit of an additional purpose available to us, and it contributes all the more to our freedom the more important this purpose is.

Carter argues that this "value-based" approach collapses into the thesis that only specific freedoms matter, and that these are the only freedoms we care about making available. His argument builds on a previous argument due to Kymlicka. In brief, the value-based approach, as Kymlicka and Carter interpret it, just says that if the freedom to do x is important to us, given our interests, then having the option to do x increases one's overall freedom: because freedom is valuable and therefore we have an interest in providing everyone with the largest possible share of freedom, consistent with everyone else's having an equally large share, this gives us reason to try and make option x available to each person, consistent with everyone else's right to have x. The problem, Kymlicka notes, is that if the freedom to do x is important, then we already have a compelling reason to make this freedom available to everyone; we can altogether dispense with this talk of overall freedom, as it plays no role in justifying the sorts of freedom expansions that the value-based approach licenses. But this, Carter argues, is just to deny that freedom is non-specifically valuable: if we only have reason to value additional options because these options themselves are valuable, then freedom as such cannot be valuable. But since freedom as such *is* valuable, this is a *reductio* against the value-based approach.

There is a gap in this argument. Carter explicitly sets out to debunk the thesis that an additional option must be "sufficiently dissimilar" from one's existing options to be freedom-expanding; but his and Kymlicka's argument seem directed against the view that an additional option must be *valuable* to count as freedom-expanding. I'm happy to grant that measures of freedom which directly reward opportunity sets for providing individuals with better options and better opportunities to live good lives do confuse the specific value of freedom with its nonspecific value (I'll argue as much in chapter 4). But an argument against this sort of measure does no work against one that rewards dissimilarity. Certainly, we can reach for measures of dissimilarity that aren't in themselves measures of value: you might, for instance, consider that two options are more dissimilar from one another the more likely it is that any given individual would fail to be indifferent between them, or the greater the expected intensity of someone's preference for one over the other. Neither of these measures tell us anything about how good the options being considered are, and they make for good measures of the intrinsic dissimilarity of two options, insofar as we should expect pairs of options to elicit more polarized value judgements the more intrinsically dissimilar they are.

I think it is now safe to conclude that the rejection of Carter's empirical approach does not force us back into the arms of the specific-freedom thesis. Appeals to dissimilarity are not hopelessly value-laden, and the need for a measure of freedom is as pressing as ever. We can therefore rest at ease in following the axiomatic approach.

Summary

Freedom matters in many ways, and we have strong intuitions that some people are freer than others – indeed, that some peoples are freer than others. The need for a measure of freedom, for a way of making our everyday commonsense comparative judgements meaningful and testable for truth. Carter's empirical approach, emerging partly as a response to the sorts of objections that Dworkin, Kristjánsson, Oppenheim, and others raised against the possibility of measuring freedom, offers one way of meeting this need. The axiomatic approach offers another, and, I argue, a better way.

Methodologically, the virtues are all on one side here: the axiomatic approach is more flexible, its assumptions are more precisely stated, more easily examined and criticized, and its focus is on the most salient question, namely, what constraints on the "offers at least as much freedom as" relation offer the best interpretation of the relation between one's available options and the degree of freedom they provide us with. The empirical approach, by contrast, largely ignores this question, and instead yields metaphysics a central place in the analysis; this leads Steiner and Carter to draw what seem to be hasty conclusions regarding the uniqueness and cardinality of their proposed measure, and ends up straightjacketing Carter into denying the relevance to the measurement of freedom of the diversity of options open to individuals. Not only is it implausible to deny that this is relevant, but to deny it as a matter of definition (since it is Carter's definition of an option that rules such concerns out) seems particularly problematic. And since the axiomatic approach survives unscathed what I take to be Carter's main criticisms of any non-empirical approach, this completes the case for the axiomatic approach.

Of course, none of this is to say that the axiomatic approach and all the assumptions that underlie it are beyond reproach, or that philosophy has no distinct contribution to make to its development. Indeed, the focus of the next chapter will precisely be to challenge one of the most central assumption of the axiomatic approach, on the basis of philosophical considerations on weakness of will.

Chapter 2

Weakness of Will and the Measurement of Freedom

Weakness of will often seems to get in the way of freedom of choice. Many of us have had the experience of being presented with an option so tempting that it felt impossible to resist, and this despite having resolved to do otherwise, or despite our better judgement. Of course, temptation is rarely so strong that it simply cannot be resisted. Nonetheless, in many circumstances, temptation makes it much *harder* to choose certain courses of actions over others. And in such circumstances, there may be uncertainty as to whether we will succeed in carrying our intentions through, if we resolve to do so. The question then arises: how does one factor in weakness of will when assessing how much freedom of choice agents are afforded by given sets of alternatives?

Standard approaches to the measurement of freedom are surprising poorly equipped to answer this question. Numerous philosophers have of course argued, persuasively, that motivational foibles like weakness of will are fetters on human freedom, most notably Charles Taylor (1985). However, the implications of this insight have largely gone unnoticed among those theorists who are interested in the *measurement* of freedom. In particular, although much has been written on how preference orderings over given sets of options and the degree of dissimilarity between these options affect the amount of freedom of choice that they offer, the thought that different options that are available might be more or less *accessible* has received little attention in the measurement literature. Indeed, as I will show below, existing approaches by and large simply rule out the possibility of unequally accessible options. As a result, these approaches overestimate how much freedom of choice certain sets of alternatives offer. My concern in this chapter will be to outline a new approach to the measurement of freedom that takes seriously differential option accessibility due to weakness of will.

I proceed as follows. In section 2, I review some of the more influential approaches to the measurement of freedom of choice. These approaches share a common feature: the way in which sets of alternatives are defined implies that they admit of no degrees in accessibility other than fully or non-accessible, and this will imply that two agents with the same nominal set of options are equally free. In section 3, I turn to the problem of weakness of will. I explore a series of examples in which the strength of an agent's will makes a crucial difference to how free they are. This turns out to be a sore point for existing approaches to the measurement of freedom, because they are unable to take account of the impact of strength of will on freedom. In section 4, I discuss and rebut three objections to my claim that weakness of will reduces freedom by making options less accessible. The first alleges that "internal" constraints like weakness of will are not relevant to the measurement of freedom. The second that the effect of weakness of will is to make our options worse, not less accessible. The third that the effect of weakness of will is to make certain courses of action unavailable to us, not merely less accessible. In section 5, I propose a new way of representing the set of alternatives that is available to agents, which incorporates a measure of the degree of accessibility of each option. Finally, in section 6, I show how my framework may be used in particular cases by applying it to issues regarding the ethics of nudge.

2.1 Standard Approaches to Measuring Freedom

Following my arguments in the previous chapter, I will be working within the axiomatic tradition in this chapter, though all the critical points that I raise here generalize to the so-called "empirical approaches" as well. Thus we begin with a finite set X of options, a set Z of opportunity sets, and we define $\succsim$, a transitive, nonsymmetric, and reflexive binary relation on Z, such that for any two opportunity sets A and B, $A \succsim B$ is interpreted as "A offers at least as much freedom as B." Our aim is to investigate what constraints to impose on $\succsim$.

It is important to underscore that all the options in X are mutually exclusive: two options are mutually exclusive if they cannot be performed in conjunction, if the individual cannot choose to do both. When we say that individuals lack free speech under totalitarian regimes, we do not mean to imply that there are specific utterances that they are prevented from making, but rather that they cannot, say, criticize the government *and then also* go on to pursue their life as normal. What makes them less free than citizens of liberal democracy is that certain conjunctively possible courses of action that are possible for free citizens are not possible for them. Likewise, if you are being mugged at gunpoint, and have to choose whether to give up your money or your life, the one thing you cannot do is walk away with your life *and* your money. Accordingly, numerous theorists (e.g., Carter, 1999, chapter 7; Pattanaik and Xu, 2015) have stressed that options must be understood as conjunctively possible courses of actions that an individual can complete in sequence, and of course any two different sequences of conjunctively possible actions define two different options.

Many of the most influential measures of freedom developed within the axiomatic approach were proposed by Pattanaik and Xu (1990), Pattanaik and Xu (1998), Pattanaik and Xu (2000), Pattanaik and Xu (2008), and Pattanaik and Xu (2015). The different ranking rules they propose illustrate the many ways one can use this framework to measure freedom of choice. In their original (1990) model, for example, they propose three restrictions on the "offers at least as much freedom as" relation. First, anyone who only has one option has no freedom of choice. Second, one always has more freedom of choice with two options than just one option. This is a very weak version of the fairly natural intuition that more options means more freedom of choice. Third, giving two people an identical additional option does not alter how much freedom they have relative to one another: if John is freer than Smith to begin with, then John remains freer than Smith if we give them both the option (which they previously lacked) to go skydiving. Each of these three assumptions seems initially plausible, but the only ranking rule over Z which satisfies all three is the "simple cardinality rule:" for any two opportunity sets A and B, $A \succsim B$ if and only if A contains at least as many elements as B.

As Pattanaik and Xu themselves point out, this rule is na´ive – indeed, they present it as a kind of impossibility result. Nevertheless, this rule provides a very natural starting point for measures of freedom, insofar as having more freedom of choice seems to be a matter of having more choices (of a suitable kind, perhaps), and indeed this rule has been the prototype for every rule that has followed. Crucially for my purposes, it has a key feature in common with every other model that has been proposed since: for any possible option x and any opportunity set A, either x is a member of A, or x is not a member of A – opportunity sets are defined as *classic* sets, and set membership is bivalent. Of course, the bivalence of set membership implies that if you and I can choose from the exact same options, then we are equally free; this, I will argue, is problematic.

Newer approaches have of course been proposed which improve significantly on the simple cardinality rule. For example, Pattanaik and Xu (1998) note that the simple cardinality rule seems objectionably insensitive to the quality of the options one has to choose from. Here they explicitly follow Sen (1993: 531) who suggests that if you already have interesting life options that are open to you, your freedom of choice does not increase if you are given the additional option of being beheaded at dawn. Likewise, if you are at a car dealership and are hesitating between several options, Sen claims your freedom does not expand if you are now given the option of buying a car that is identical to one already on offer, except for a defective gearbox. The lesson Sen draws from this is that an individual's preferences over the options that are available to them matter to how much freedom those options offer them. In consequence, numerous authors have proposed to impose some sort of minimal desirability criterion on opportunity sets. Pattanaik and Xu (1998, p. 182), for instance, propose that an additional option should only be counted as freedom-expanding if and only if it is such that at least some rational persons could prefer it to its available alternatives (see also Sudgen and Jones, 1982; Parfit, 1981, chapter 1, footnote 39; Puppe, 1996; and Foster, 2011 for further discussion). This restriction looks like a clear improvement over the previous rule, but it does not change the fundamentally bivalent character of opportunity set membership.

This feature remains even when ranking principles are refined in other directions. For instance, numerous authors have emphasized that one's degree of freedom depends crucially on how diverse one's options are. This point has been emphasized by Charles Taylor (1985), who argues that freedom only matters to us because we are purposive beings, and so one should only count as freer than another if one is free to pursue a greater range of purposes, and if one is free to pursue more important purposes (see also Arneson, 1985 and Raz, 1988, pp. 408-409). Accordingly, various authors have proposed ways of ranking sets according to how diverse they are. The simplest such proposal comes from Pattanaik and Xu (2000), who essentially propose to rank opportunity sets by how many types of options they contain, where an option type is a set of available options that are all similar to one another (see also Pattanaik and Xu, 2008; Bervoets and Gravel, 2007; Hees, 2004 for further interpretations). This approach again seems to improve upon the simple cardinality rule, but bivalence remains an implicit assumption. Things are not different, incidentally, under the "empirical approach" defended by Carter (1999) and Kramer, 2003, where bivalence is in fact "promoted" to an *ex*plicit assumption.

The key problem with the general approach in this literature was anticipated by Berlin (1969, p. 130), when he claimed that the extent of an individual's freedom depends not only on "how many possibilities are open to, them;" but also on "how easy or difficult each of these possibilities is to actualize," and "how far they are closed and open by deliberate human acts." One way of cashing our Berlin's insight is to say that options may be more-or-less *accessible* to individuals, even as they are *available* to them. The problem that this lesson poses for the traditional models comes out clearly when we consider cases of weakness of will.

2.2 The Problem Posed by Weakness of Will

2.2.1 What is Weakness of Will?

We all frequently confront temptation. When we succumb to it contrary to our better judgement, then, according to the orthodox view on the subject, we display weakness of will (e.g., M. Smith, 2003; Pettit, 2003; Stroud, 2003). More precisely, according to Davidson's (1969) seminal analysis, weakness of will is an intentional action, motivated by some strong present urge, yet contrary to the individual's all-thing-considered better judgement. A standard case of weakness of will, on this account, is that of a dieter, who judges that she ought to stick to her diet, but nonetheless splurges on cream puffs and cheese cake. Or a smoker who judges that they ought to quit, but still reaches for the next cigarette.

Discussions on the impact of weakness of will on individual freedom generally presuppose this orthodox view of weakness of will. However, as I will argue later, this makes the resulting analyses undesirably narrow, as there are other plausible views on the table. Richard Holton and Berridge. (2013, p. 77), for example, has articulated a heterodox but nonetheless influential view according to which weakness of will really consists in over-readily or irrationally reconsidering one's resolutions, where a resolution is understood as a future-directed intention which is formed precisely for the purpose of overcoming one's anticipated future inclinations to act contrary to one's intention. On this view, it is irrational to revise one's resolutions if we do so under the pressure of just the urges which our resolution was formed to defeat, and no new decision-relevant information has come to light.

Holton's view is partly motivated by the fact that many of what we consider to be paradigmatic cases of weakness of will are typically preceeded by what Holton (2006, p. 98) calls "judgment shift": those who give in to their urges and break their resolutions typically first come to judge that it is better, after all, to do so. A common example: you swear to quit smoking, hold fast for a while, then an oblivious coworker offers you a smoke, and at first you resist, but then say to yourself: "ah, well, why not just one? One cigarette won't kill me. In fact, I'd enjoy it, and I can still quit tomorrow!" Then all resistance fades, and you cave to the urge to smoke. This smoker clearly displays weakness of will, and yet he doesn't appear to be acting against his better judgment.

To avoid committing myself in advance to any particular view, I will consider "weakness of will" to be a cluster concept, covering a range of irrational motivational foibles. All I will assume is what empirical psychology tells us: first, that weakness of will (or "failure of self-control," among psychologists, e.g., Fudenberg and Levine, 2012) manifests as a conflict between some present urge and some antecedent motivation (e.g., better judgment, resolutions, etc.), and which may hinder our ability to act on certain intentions. Second, that it is typically cue-driven, in that these urges usually don't form spontaneously, but rather in response to environmental stimuli (Holton and Berridge., 2013). Addictive desires are the most vivid case in point: addicts become sensitized to certain cues associated with past drug use, and subsequent exposure to those cues immediately results in very strong motivational responses, even in former addicts who have not used for a long time (see Robinson and Berridge, 2003). This is why addicts who have long been sober are much more likely to use drugs in the future than individuals who have never used drugs before. Also, more salient cues typically trigger stronger urges (Holton, 2006, p. 97).

By contrast, strength of will – the capacity to recruit the motivation to defeat one's present urges – is generally thought to be enabled by the use of a distinctive cognitive capacity, willpower, which is effortful to employ but enables us to enforce our better judgment and commitments in the face of contrary desires (Holton, 2006, p. 112).

There are two dominant models in the experimental psychology literature on how this faculty operates. The "cognitive depletion" model of willpower, pioneered by Muraven, Diane, and Roy F. Baumeister (1998), interprets willpower as a kind of cognitive resource which we call upon at will to recruit the motivation to carry through our resolutions. This resource is limited in supply over the short term, and "spent" in resisting temptation (this is felt as the exertion of effort), so that previous use of our willpower leaves less available for subsequent use in the short term. Individuals can thus progressively exhaust their willpower, making future attempts at resoluteness less likely to succeed. The phenomenon of "ego depletion" is often cited as providing evidence for this view. To cite just one example, the experiments of Muraven, Diane, and Roy F. Baumeister (1998) seemed to show that forcing oneself to resist temptation in one period makes one less likely to persist in one's efforts to complete a difficult or frustrating but completely unrelated task in the next period. In their initial study, forcing oneself to eat radishes instead of the more tempting chocolates made one less likely, in the subsequent period of the experiment, to persist in one's efforts to solve puzzles. This was taken to show that willpower is depleted with use.

This is not to say that every case of abandoning one's resolutions is a case of having exhausted one's willpower; it is surely quite common for people to give into temptation without ever calling upon their willpower to resist the urge to yield ("I know I resolved not to eat chocolate cake, but things are different today: I'm on vacation!"). As Rizzo (2016, p. 22) puts it, willpower is just another scarce resource: we can choose how to allocate it across different consumption bundles, but there is a hard cap on how much we can buy. Importantly, individuals can become better at exercising their willpower. According to Vohs and R. Baumeister (2004), the more one succeeds in being resolute, and the more one cultivates certain habits of plan-making, the easier it becomes to recruit motivation going forward (if one is keeping the resource metaphor in mind: training can give you more motivation-recruiting-bang for your effort-of-will-buck). This interpretation has many adherents in behavioural economics (see for example Rizzo, 2016; Vohs and Heatherton, 2000; Fudenberg and Levine, 2012), and was until recently the orthodox view.

In recent years, however, the "attentional myopia" model of willpower has emerged as a major rival theory. On this interpretation, weakness of will arises in situations when our attention narrows on the most salient cues (due to cognitive stress or prior sensitization), which suggest resolution-violating behaviours (see Mann and Ward, 2007, p. 280 and Michael Inzlicht and Schmeichel, 2012, p. 452). Cognitive load makes us susceptible to temptation, on this view, not because complex cognition consumes willpower, but because the way we typically deal with cognitively demanding situations is by focusing our attention on the most salient features of that situation (Mann and Ward, 2007, p. 281). So for example, one experiment conducted by Mann and Ward (2007, pp. 281-282) found that dieters who performed cognitively demanding tasks consumed twice as much milkshake right afterwards when placed in a room containing salient food items than when in a room containing conspicuously placed scales and dieting books, whereas dieters who performed cognitively low-load tasks were significantly less influenced by cues.

Strength of will is interpreted on this view not as a cognitive resource, but as a *skill* in refocusing one's attention. Elkins-Brown, Teper, and M Inzlicht (2017) argue that it is possible to recruit the additional motivation to maintain one's resolutions in the face of temptation by refocusing one's attention on cues which prompt self-control (e.g., to think about how the action would be socially inappropriate), in particular through exercises of mindfulness, which make one more receptive the cues which would otherwise lack salience. It's been found, for example, that those who practice meditation experience better outcomes in smoking cessation (see Elwafi et al., 2013; Y. Y. Tang, R. Tang, and Posner, 2013) and alcohol use (see Bowen et al., 2006) than control groups, and there is some evidence to suggest meditation may be an effective treatment for drug abuse disorders Chiesa and Serretti, 2014. Elkins-Brown, Teper, and M Inzlicht (2017, p. 5) argue this is because "mindfulness meditators are better prepared to acknowledge moment-to-moment affect that signals the need for self-control."

There is substantial evidence for both dominant views, but, regardless of which model has got the details right, there are some broad-stroke conclusions that come from the experimental psychology literature. First willpower is a cognitive resource or skill which we can call upon to recruit the motivation necessary to defeat resolution-inconsistent urges. Second, some agents are better than others at leveraging this faculty. This is well attested to by the fact that individuals can train themselves to better resist temptation, for example by practicing mindfulness exercises, or cultivating habits of plan-making. And, third, cognitively demanding tasks and exposure to temptation temporarily undermine one's ability to subsequently recruit additional urge-defeating motivation (see Muraven, Diane, and Roy F. Baumeister, 1998; Mann and Ward, 2007; Holton, 2006).

2.2.2 Problem Cases

Let us now look at some cases. Consider first Weak-willed Willy and Resolute Regina. Suppose that every night for a whole week, both have the choice between four alternatives: completing job applications, solving logic problem sets, working on their essays, and going out to drink with friends. In other words, their opportunity sets are identical. Both would love nothing more than to go out with their friends, but have resolved not to because this would adversely affect their capacity to meet urgent deadlines with respect to their work, and they both judge that meeting these deadlines is more important than enjoying themselves that week. Regina, being resolute, is pretty consistently successful in recruiting the motivation to resist temptation, and she is very skilled at shifting her attention towards cues that facilitate self-control. She is quite reliable: if she says she'll do something, or that she judges that she ought to do it, it's generally a sure bet that she will. Willy, by contrast, is a flake, generally unsuccessful in resisting temptation when it presents itself, and unskilled at refocusing his attention. He is almost always carried away by his present urges to renege on his previously stated intentions and judgements.

The first thing to observe here is that on standard constructions of opportunity sets, Willy and Regina are equally free. In every relevant respect, after all, their opportunity sets are the same: the options in both sets are equally many and diverse, they are all desirable, and present Willy and Regina with the exact same sorts of opportunities. There's nothing one can do that the other cannot (it may be more or less difficult, or more or less pleasant, but they both can do it). Thus, on the traditional picture, they simply cannot be unequally free. But this looks like a mistake. Indeed, whatever the outcome of this case, I submit that Willy is less free than Regina: either he expends his willpower reserves more quickly, less efficiently than she does in attempting to be resolute, or he is simply far less skilled at shifting his attention to recruit the motivation to defeat his urges. Sure, he may well succeed in standing firm in the face of temptation, but there is a clear sense in which he is much less *able* than she of doing so.

But in what sense does this make Willy less free? Charles Taylor (1985, p. 220) argues that we "experience our desires and purposes as qualitatively discriminated, as higher or lower, noble or base, integrated or fragmented, significant or trivial, good and bad." And what it is for an end to be qualitatively superior (inferior), on his view, is for it to be the sort of end that we desire to (not) desire. Of course, as he observes, what we actually desire and what we desire to desire can come apart, thus we may upon occasion pursue base goals instead of ones we deem worthier – this is just what it is to be weak-willed. But to be driven by one's defects to pursue base ends is *ipso facto* to be constrained in one's ability to pursue the ends one truly finds worthwhile, and this is injurious to our freedom, Taylor insists.[1]

Taylor seems to me to have gotten something right here: motivational defects like weakness of will constrain freedom because they make us less capable of pursuing certain aims. That said, this on its own remains quite vague (what is it for a person to be more-or-less capable, for instance?), and in some respects Taylor's analysis lacks generality. For one thing, it is very much tied to the orthodox view of weakness of will, and cannot be extended to non-standard views. If, as Holton argues, weakness of will is often characterized by the reevaluation of what we are initially disposed to regard as base desires into acceptable, even worthy ones, then it cannot be in those cases that our freedom is constrained because we are unable to pursue ends we find worthwhile. Rather, in such cases, it is our ability to carry out our resolutions that is impaired.

This point bleeds into a further one. For Taylor, weakness of will is a restriction on freedom only because it makes us choose *lesser* options over *better* ones; this analysis might seem too value-laden. It suggests a view on which something only constrains our freedom if it make us worse off, in some way. Isn't the problem more generally just that susceptibility to temptation undermines one's ability to act on certain aims and intentions, regardless of whether those are ones we strongly

[1] There are other influential views one could consider on how internal constraints impair individual freedom. Crocker (1980, pp. 36-43) and Christman (2001), for instance, both argue that heteronomy impairs freedom, and that the less autonomous one's decision to ϕ, the less free one was not to ϕ. Taylor's analysis is particularly salient for us, however, because his account is explicitly concerned with explaining how *weakness of will* constrains individual freedom, unlike Crocker's or Christman's. And indeed, their analyses have no obvious application to the present case. The paradigm cases of heteronomy Crocker considers are cases of impaired consciousness and coercion, while for Christman the autonomous character of a decision depends on the desires driving the decision being formed by a process the individual would not have resisted had they attended to it; by either author's lights, Willy and Regina might both count as fully autonomous.

identify with? A workaholic might desire to desire to work; nonetheless, workaholic compulsions plausibly impair one's freedom to relax and go on vacation. A more value-neutral analysis of constraints is needed to account for these kinds of cases, one which allows constraints to be either good or bad, and only requires that they make one less *capable* (we will return to the issue of value-neutrality in section 4.2).

In light of the above considerations, I propose the following analysis of Willy and Regina's situation. Willy is less free than Regina because although all four options *are* available to both of them, Willy's work options are less *accessible* to him, in the sense that, *conditional on his intending to work*, he is less likely to successfully enforce his intention and *actually* work than Regina is likely to, conditional on her intending to work. It is as though, in forming the intention to work, Willy and Regina both take out a lottery ticket which gives them a particular chance of successfully working, and a particular chance of failing and choosing to go out instead; because Willy is weak-willed, his ticket gives him longer odds of success. This analysis is more precise than Taylor's, as it cashes out in terms of conditional probabilities the sense in which Willy is *less capable* than Regina of enforcing his practical attitudes. But it is also more general, as it is plainly value-neutral,[2] and does not assume a particular picture of weakness of will, only that it is a foible which may hinder our ability to act on certain intentions, but which may be overcome by an effort of willpower, and which some are more adept at overcoming.

This analysis also happens to be congruent with the sort of analysis which

[2] Carter (1999, pp. 153-155) discusses a proposal very similar to mine, on which we must assign numerals between 0 and 1 to each option we are externally unconstrained from performing, based on the degree of internal constraint, with "1" representing complete internal freedom to perform it, and "0" the complete internal unfreedom to perform it. Carter dismisses this proposal on the grounds that the only way to distinguish internally constraining and non-constraining desires is by reference to the value of the purposes that these desires hinder. Since I defend a version of this view, I feel I must make a couple points in response: first, it is unclear in Carter's treatment how the numbers are to be interpreted. He is led to consider this view by the thought that we ought to measure "the internal unfreedom of an agent to do x in terms of her propensity to do some other thing," which invites a probabilistic interpretation of the numbers. However, he goes on to argue that the degrees of internal constraint represented by the numbers "would appear to correspond to the 'strength of will' that is required to overcome them," and he quotes Flathman (1987) taxonomy of involuntary behaviours, ordered by their degree of involuntariness, as offering a possible measure of the degree of internal constraint on an individual. This suggestion is nonprobabilistic, and leaves mysterious how one arrives at a cardinal measure of the strength of efforts of will. My approach has the virtue of clarity: internal constraints have an unambiguous interpretation in terms conditional probabilities. Second, and more importantly, it is false that a notion of internal constraint must be objectionably value-laden: my own proposal makes no reference to values or preferences, only intentions and conditional probabilities. Carter is therefore too quick to reject this approach.

Manley and Wasserman (2008) argue ought to be given of all dispositional properties. A dispositional property is the property of behaving in a certain way if certain conditions obtain. More precisely, Manley and Wasserman (2008, p. 76) argue that an object is disposed to behave in a certain way when a given stimulus condition obtains if and only if the object would behave in that way in suitably many cases where the stimulus condition obtains. And accordingly, one object is more disposed to behave in a certain way than another if and only if the first object would exhibit the behaviour of interest in more cases where the relevant stimulus condition obtains than the second object would. Importantly, some dispositions require no specific stimulus conditions: if one is prone to anger in any situation whatever, one is irascible. Being strong willed is arguably a similar sort of disposition: it is to be reliably disposed to act on one's intentions (whatever those are) in suitably many cases, notably in cases of temptation. It would follow, on Manley and Wasserman's analysis, that to be more-or-less strong-willed is just to be more-or-less likely to successfully enforce one's intentions, whatever those are. This is just what I've argued. Thus a general analysis of dispositions supports my view of why weakness of will is a constraint upon freedom: it depresses one's chances of enforcing one's intentions.

Note that we can extend my analysis to other cases. Consider a person suffering from intense depression who is trying to hide her condition. Such a person can, usually, muster the will to go to work, attend social events, go out with her friends, and so on; it can feel like drowning, but she may succeed in hiding her condition if she's determined. Sometimes, however, her will fails, and she will wake up to find she can't get out of bed. She may believe that the best thing for her right now is to try and maintain a "normal" social life, and she may know that staying in bed all day will only make her feel worse about herself, but still she fails to recruit the motivation to unmoor herself from her bedposts. Our patient is arguably not weak-willed – she is not driven by some strong desire to act against her better judgment or to break her resolutions – but is rather prey to what Aquinas called accidie, or the total loss of motivation. Nonetheless, she would rightly view the end of her depression as a liberation, and for the same reason as before: not because it affords her any more options, but because it restores to her control over her choices. Once it is no longer a battle of uncertain outcome to recruit the motivation to get out of bed, to go to work, or to make dinner, she can once again count on herself to act on her practical attitudes (i.e., intentions, resolutions, better judgments, etc.).

Going further, my analysis suggests Willy might have more freedom if his friends were out of town, leaving him with only the three work options. Free from temptation, free from the constant, painful battle to maintain his resolution which he wasn't guaranteed to successfully carry out, he now has no trouble at all setting priorities between work tasks and delivering quality work on time for all his deadlines. The loss of his tempting option may therefore leave Willy with more freedom, by making his remaining options more accessible. This is difficult to make sense of under standard approaches to the measurement of freedom: there's nothing that Willy can do now that he couldn't do before, no outcome that he can achieve now that he couldn't achieve before – much to the contrary: there's less he can do now – so it looks very much as though we've simply removed an option from him, and standard monotonicity assumptions imply that this *never* increases one's freedom.[3]

We can draw some preliminary lessons here. The degree of freedom offered by one's opportunity set depends, first, on whether choosing any of the options on offer requires one to recruit any motivation to defeat one's urges to choose otherwise; second, on how strong-willed the agent is, that is, how capable they are of recruiting motivation; and third, on what other options are available. A fourth important factor is the salience of one's options. This is brought out clearly in the case of addiction.

Consider Perseus and Cassandra, two cocaine addicts who successfully complete their detox program and return to their homes. They both resolve never to do drugs again, they are equally strong-willed, and face the same constraints, so that the same options are available to them. Anything the one can do, the other can as well. We can even assume that they share the same preferences and value the same alternatives. On any standard measure of freedom, they are equally free. The only difference between the two is that while Perseus still lives in his old flat where he is regularly exposed to drug cues, Cassandra has the foresight to move to a part of town where she does not face any such cues (though it's a short drive to a dealer). In other words, Cassandra and Perseus have the same (or very nearly the same) opportunity set, yet I contend that Cassandra will have more freedom of choice when she gets home from detox than Perseus.

Indeed, given that Perseus is exposed to various drug cues, the possibility of doing cocaine will be very salient to him, triggering his addictive desires. Plainly, the attentional myopia model of willpower predicts that his attention will narrow on the salient behavioural cues in his environment, which are very suggestive of cocaine, especially if he is ever under conditions of cognitive stress. And on resource models of willpower, resisting the pull of his addictive desires will very seriously

[3] Pattanaik and Xu (1998) are an exception here; if an option dominates every other option in a set under every rational preference ordering, then removing that option might increase one's freedom. Clearly though, this is not the situation here, so their monotonicity assumptions would likewise imply that that removing Willy's fourth does not increase his freedom.

deplete his willpower reserves, requiring a great deal of effort. On both models then, resisting his addictive desires is extremely difficult, and he is not all that likely to succeed anyway. In contrast, since Cassandra is not exposed to these cues, her addictive desires are not triggered, and so she will have no difficulty in maintaining her resolution to stay clean. Hence, despite the fact that Perseus and Cassandra have identical opportunity sets and are equally strong willed, the simple fact that the "wrong" option is made salient to Perseus and not to Cassandra means that the former has less freedom of choice, and for the same reason as before: because of his addictive desires, he is less capable of making certain choices, and so some of his options are less than fully accessible to him.

The point that these examples have laboured to bring out is simply this: options admit of degrees in accessibility, and what it is for an option to be more-or-less accessible is simply for it to be the case that, conditional upon your intending to perform it, you are more-or-less likely to perform it, successfully enforcing your intention. Motivational defects like weakness of will clearly inject great heterogeneity in people's capability to make choices, and standard approaches to the measurement of freedom are oblivious to these distinctions. To more meaningfully compare individual freedom requires a framework that captures the heterogeneity in the capability of different agents to do, be, and have things worth wanting.

2.3 Objections

Before going on to propose a new model, I consider three responses to the argument I've put just put forward. Some will be skeptical that weakness of will really does impose any constraint upon individual freedom. Others might concede that weakness of will is a constraint, but argue that this is either because weakness of will somehow impairs the value of our choices, or because it actually rules out altogether certain conjunctive courses of action. Let us look at each in turn.

2.3.1 Negative freedom

Recall that proponents of negative conceptions of freedom, such as Oppenheim (1981), Miller (1983), Steiner (1994), and Carter (1999) and Kramer (2003) claim that one is free to ϕ if and only if one is not prevented from ϕ-ing by another (through deceit, coercion, etc.). There is a very sharp distinction on such views between being *unable* to do something and being *unfree* to do it. Nature constrains us in all sorts of ways, but does not thereby make us want for freedom. Paraplegics are not unfree to climb mountains, though they cannot do so. Accordingly, a first response to my argument might be that Willy and Perseus are not in any way less free than their female counterparts, nor does depression constrain one's freedom, because weakness of will is a feature of one's psychology, not a barrier to choice imposed on us by others.

As stated, this claim is surely too strongly put. True, in my previous examples, weakness of will seems to act as an internal constraint. But weakness of will can arise through many different processes, some of which have their origin in the deliberate behaviour of others, and in those cases it looks much more like an external constraint. Suppose I tempt an addict by waving bags of cocaine under their nose, thereby triggering their addictive desires and making it exceedingly difficult for them to go on about their business as they had planned; is it I, or defective brain chemistry, that constrains our addict's ability to stay clean? Or suppose I am a terrible boss who creates a hostile, oppressive work environment, leading some of my employees to become so depressed that they miss days of work; is it I, or my employee's own unfortunate but natural tendency to depression, that now binds them to their beds?

In cases like these, I act like a polluter: I am introducing temptations and sources of cognitive stress into the social environment, triggering difficultly resisted compulsions in others who, but for my behaviour, would be considerably more capable to do what they wanted to. This is not so different from how, by poisoning the soil with heavy metals, industrial waste dumping impedes our freedom to build houses, grow crops, raise farm animals, or procure drinking water. These acts of pollution clearly interfere with the decisions of others, and constrain what they may do.

In fact, we can quickly construct much more chilling scenarios than ones involving pollution. Imagine, for instance, that a megacorporation uses subliminal messaging or highly aggressive advertising and propaganda to manufacture addiction-like desires in us, which it then triggers at its convenience to manipulate our behaviour, in contravention to our better judgements and resolutions. These desires may well be resistible, with enough strength of will, but it is clear that the corporation is interfering with our choices, curtailing our freedom. But what is remarkable in this case is that the corporation is controlling us by exploiting our susceptibility to weakness of will. This and the pollution scenario reveal that *even if* one insists that the only genuine constraints on freedom are those imposed on us by others, one must concede that there *are* cases in which weakness of will (and related motivational defects like depression and compulsion) *are* constraints imposed on us by others. Thus, at least in those cases, the proponent of negative freedom must care about how weakness of will constrains our freedom, and what my treatment of this question shows is that they must give up bivalence.

Finally, it isn't clear that the distinction between internal constraints and external constraints is well-formed. Suppose I fall into a pit trap and cannot get out. My freedom is curtailed in this case, but is this because I face the *external* constraint that the pit was dug too deep, or because I face the *internal* constraint that I cannot jump high enough? Likewise, is the constraint Willy faces purely an internal one? He only struggles to be resolute, after all, because he's tempted *by his friends' decision* to go out to the pub. It looks as though whenever a person is unfree to do something, this is always due to the joint impact of internal and external factors. Thus external factors, in particular, cannot on their own prevent anyone from doing anything; they require the cooperation of internal factors. This makes it difficult to articulate a coherent picture on which one is free to ϕ just in case there are no "external constraints" on one's ϕ-ing. For this reason, I submit it is better to side with Sen, and accept that you are free to ϕ just in case you can, in fact, ϕ.

This isn't to deny that there is a morally salient difference between the situation of a person who is barred from attending university by the State because she is a woman, and the situation of a person who simply can't afford tuition. But the difference, I think, is that one is a *worse* form of unfreedom than the other, not that one is a case of unfreedom and the other a case of something else. Likewise, to take an example from Kramer (2003, p. 367), it is much *worse* to be unable to walk more than a hundred meters because one has been assaulted than to be unable to walk more than a hundred meters because of a disability. There *is* a relevant contrast between these two cases. But as Sen (2009, p. 306) emphasizes, there is already a morally salient contrast between being capable of doing something and being incapable of doing it. It is this second contrast which is of greatest interest to us when we aim at a measure of freedom: we want to know how *capable* people are. And with this in mind, the problem posed by weakness of will becomes particularly salient, as weakness of will does constrain individuals' capabilities, however it arises.

2.3.2 The value of options

Of course, one might accept everything I've just said, but contest my explanation of how weakness of will constrains individual freedom. Numerous theorists have defended what Carter (1999, p. 170) refers to as a "value-based" account of freedom, according which one's degree of freedom is determined, in some robust way, by the value of one's options; accordingly, a critic might object that the real reason weakness of will constrains individual freedom is that it somehow impairs the value of one's opportunity set.

Sen (1990c), Arneson (1985), and Crocker (1980), and others have argued that, all else being equal, an individual has more freedom of choice than another if their options are better. "Better" can be interpreted either subjectively, as reflecting the individual's own preferences or value judgments, or objectively, as reflecting some impartial assessment of how good the options are. As we saw earlier, Taylor deems you freer than I if you can pursue the aims you desire to desire and I cannot, so his is a subjective interpretation. Sen (1990), similarly, argues that if you and I both only have one option, but you like the one you have and I do not like mine, then you must be seen to enjoy a greater degree of freedom. In contrast, Raz (1988, p. 409) takes a more objective view, arguing that ("positive") freedom is expanded by whatever expands autonomy, which is expanded in proportion to (among other things) the diversity of options which one has and which allow one to develop all of one's mental and physical abilities Raz (1988, p. 376).

In any case, if one is going to rely on preferences or value judgments to (partially) determine the degree of freedom offered by a set of options, then it becomes necessary to specify options in terms of all the features that are relevant to our assessment of these options, and here it might be argued that the way in which I specified the options in section 3 was incomplete. I alleged that Willy and Regina both have the same work options, but in fact (so the objection goes) the options available to Regina are: "work on problem sets without exerting willpower," "work on essay without exerting willpower", etc., while the options available to Willy: "struggle to overcome temptation and work on problem sets", "struggle to overcome temptation and work on essay," etc. Accordingly, when Willy's friends leave town, it's not that options which were previously available became more accessible, it's rather that the options "struggle to x," "struggle to y", "struggle to z" were replaced with the options "do x without struggle," "do y without struggle," "do z without struggle."

Redescription on its own doesn't get us very far, since Willy and Regina still have equally many options to choose from,[4] which seem equally diverse, and it

[4] This may seem surprising at first glance: Willy cannot do x, y, or z without struggling, whereas Regina can, so doesn't it follow that she disposes of more options than Willy? But notice that just as Willy can't do x, y, z, without struggling, Regina can't choose to struggle-to-do-x, y, z – she can only choose to do x, y, or z without struggling, since there's nothing for her to struggle against – whereas Willy *can* choose to struggle-to-do-x, y, z. It's not as though, being strong-willed, she can choose between doing x without struggle and doing x by struggling. If it requires no effort of you to turn down a friend's offer of a cigarette, you can't make yourself have to struggle to turn it down. So both Willy and Regina have a choice of exactly three options, differing only in how strenuous an effort of will they require in order to be carried out.

seems rationally permissible for both Will and Regina to choose any of their available options. However, once options are appropriately redescribed in this way, the following explanation for why Willy is less free than Regina becomes available: it is not that his options are less accessible to him than Regina's are to her, but rather more simply that her options are preferable, insofar as they require less of a struggle. Similar explanations are available for Perseus and Cassandra, and for our depressed patient. What addiction and depression do to a person is not to make certain courses of action less accessible to them, but merely to make certain courses of action altogether impossible without struggle and torturous effort. Thus bivalence is saved: options are either available or they are not, full stop, and weakness of will only impacts our freedom by making our options more-or-less valuable.

There are two reasons why it would be a mistake to pursue this strategy. Firstly, tying the degree of an individual's freedom this tightly to the value of their options risks eliding the distinction between the amount of freedom one has, and the value of that freedom to us. This is a contrast worth preserving, because in order to have a clear-eyed view of what trade-offs we are making in weighing freedom against other values, it pays to have a measure of freedom that does not make one's degree of freedom dependent on the extent to which it promotes other values, with which it may compete.[5] The value of one's options of course matters – greatly, in fact – but having more freedom isn't primarily a matter of achieving better outcomes, it's a matter of being more *capable* of doing, having, being, and becoming things which we have (at least minimal) reason to value.

Accordingly, when confronting the case of Willy and Regina, the critical question is not which of the two has the better options, or who achieves the most desirable outcomes, but which of the two is most capable of achieving ends that one *might* have reason to value. I claim Regina to be more capable. She is the one who has the most control over what she ultimately does: if she judges that deadlines don't really matter, she can choose to go out, but if she judges that she ought to solve problem sets, or resolves to do so, she is capable of reliably enforcing her judgements and commitments. My approach gives us a way of asserting that

[5] Note that the ranking rule mentioned in section 2 according to which an option only contributes to your freedom if at least some rational person could prefer it to its available alternatives, while obviously not entirely value-neutral, nonetheless preserves this distinction. This view allows an option to count as freedom-expanding even if no one actually prefers it, and it explicitly denies that, all else being equal, having better options makes you freer: an individual with three terrific option is exactly as free as an individual with three middling options (or even three bad options), provided any of the three middling options could be rationally preferred to the other two. Accordingly, this ranking rule would fail to imply that Willy is less free than Regina (see footnote 3), unless one accepts my proposal to reject bivalence.

Regina is freer than Willy without committing ourselves to any judgement regarding the value of their options, and this is an attractive feature of the approach.

Secondly, even if we do accept the value-based view, it is at most a partial story, because options are not necessarily worse if they are more effortful. Some people may find that there is great value in having to struggle to get what they want, and that having easy choices cheapens the value of what is obtained; they may place greater value on certain options precisely because they require more willpower. Nietzsche (1968, TI, 92), for instance, reserves high praise for those individuals locked in a constant struggle to overcome themselves and their limitations:

> "That one has become more indifferent to hardship, toil, privation, even to life. The man who has become free... spurns the contemptible sort of well-being dreamed of by shopkeepers, Christians, cows, women, Englishmen and other democrats. The free man is a warrior... One would have to seek the highest type of free man where the greatest resistance is constantly being overcome."

Willy, likewise, perhaps because he is an avid student of Nietzsche, may spurn the ease with which Regina goes through life, never prey to her passions or experiencing the pain of self-struggle, and he may thus prefer his own more effortful opportunity set. Indeed, Willy may object to any paternalistic intervention that would remove temptation from his sight, arguing that it would rob him of his struggle, of his chance to triumph over himself. Still I insist that he and individuals like him are, in at least one respect, less free to choose than they would be if they were not required to recruit any willpower in order to choose any of their actions: when Willy reaches out to choose to do what he has resolved to, he is likelier to find his reach too short, and choose otherwise in the end. Whatever the impact of an option's effortfulness on that option's desirability, there is in any case a separate impact on how much freedom this leaves one with.

2.3.3 Conjunctively possible courses of action

Finally, a critic might object that my analysis misunderstands the mechanics of weakness of will, as they are revealed to us by the experimental literature. More precisely, it might be argued that weakness of will ought not be understood as a probabilistic constraint on one's ability to carry out certain courses of action conditional on intending to carry them out, but rather as a ruling out of certain classes of conjunctive options.

What the ego depletion experiments purportedly reveal is that if one uses one's willpower in one period to resist temptation, then it becomes much less likely that one will resist temptation in the next period; one possible interpretation of these findings is that one is fundamentally limited in how much motivation one can recruit over a given period of time, with some being more limited than others. So suppose I believe I ought to do x and y, but that I am in addition also very strongly tempted to do z, and that the timeframe in which these options are performable is limited in such a way that I can only perform two of x, y, and z in sequence; doing what I think I ought (and have resolved to do) means not doing what I really most want to do. If I have the will to overcome this temptation, then my opportunity set is $\{(x,y),(x,z),(y,z),(y,x),(z,x),(z,y)\}$, but if performing x and y (in any order) requires me to recruit more motivation than I am actually capable of, then in fact my opportunity set is $\{(x,z),(y,z),(z,x),(z,y)\}$.

If we now return to the Willy and Regina scenario, then, on this view, the correct explanation for why Willy is less free than Regina is that since Willy is weaker willed than Regina, there is less motivation that he is able to recruit over any given period of time than she is. His "willpower budget" over a given time horizon is smaller. And therefore, *if* he recruits the motivation to resist temptation on one night, he will have less willpower left to "spend" going forward than Regina, meaning that some possible courses of action which would be open to her (counterfactually, at least) would not be open to him. In this way, bivalence is saved: weakness doesn't make options less accessible, it just makes some options unavailable.

In reply to this objection, I would point out that the idea of a *literal* willpower budget can only be meaningful on the resource model of willpower. On the attentional myopia model, cases like Willy's and Regina's break down in purely probabilistic terms: being more-or-less skilled at recruiting motivation by refocusing one's attentions just is a matter of being more likely to act on one's resolutions and to follow one's better judgement. If the resource model turns out not be empirically adequate, then this "budget constraint" idea cannot explain why Willy appears to be less free than Regina. My proposal, however, is compatible with any substantive view of willpower. After all, it's worth reminding ourselves that whether or not they reveal the existence of a literal willpower budget, ego depletion experiments certainly suggest that depleting one's willpower reserves makes it *less likely* that one will succeed in being resolute going forward.

Perhaps it will be replied here that if one isn't guaranteed to succeed in carrying out a particular course of action, then one isn't free to carry it out at all, but only free to try to carry it out. On this proposal, the only difference between Willy and Regina is that one is likelier to carry out courses of action that they are both free to try to carry out. It would follow that Regina really isn't freer than Willy to do anything. Weakness of will can only make a difference to how free people are if it makes a difference to how many courses of action they can be guaranteed to carry out if they try.

This would be a very radical reply. Only necessary propositions have a probability 1 of being true, and there is simply no course of action which it is necessarily true that we would succeed in carrying out if we tried. It would follow, on this reply, that none of us is free to do anything, only to try to do things. But anyone can *try* to do anything: I can *try* to flap my arms and fly to the seventh moon of Jupiter, and so can you. Are we all equally free then? This would be an unfortunate conclusion. Rather than admit to this, I suggest we simply accept Berlin's view that our freedom depends both on how many possibilities are open to us, and on how easy or difficult these possibilities are to actualize – and I propose to explicate the sense in which possibilities can be easier or harder to actualize in terms of conditional probabilities, which requires giving up bivalence.

2.4 Fuzzy Freedom

I now present my view on how freedom of choice should be measured. There is actually a simple way of capturing the idea that options may be more-or-less accessible to an agent, and this is to represent opportunity sets as so-called "fuzzy sets," where the membership of an element in the set is not bivalent. I show below that this will allow us to generalize the models canvassed in section 1 so that they yield the intuitively correct rankings in the problem cases laid out in section 2.

But first, some definitions. I define an opportunity set as a pair (X, m), where X is the set of all possible options, assumed finite, and m is a membership function, which assigns a value (called a membership grade) of between 0 and 1 to every element in X. Z^* shall denote the (classic) set of all such pairs. For any set A, μ_A denotes the membership function of A, and A's cardinality (noted $|A|$) is just the sum of the membership grades of all its elements. Now, $\forall x \in X, x$ is called:

Fully accessible in the (fuzzy) set A if $\mu_A(x) = 1$

Not available in the (fuzzy) set A if $\mu_A(x) = 0$

Partially accessible in the (fuzzy) set A if $0 < \mu_A(x) < 1$

Representing opportunity sets as fuzzy sets is extremely natural as an extension of the approaches discussed in section 1. The concept of a fuzzy set is simply a generalisation of the concept of a set – classic sets are degenerate fuzzy sets in which all elements are assumed to have a membership grade of either 1 or 0. Fuzzy sets retain the concept of cardinality, and all classic set operators have fuzzy analogues; the only difference is that the device of fuzzy sets allows us to explicitly represent the idea that options admit of degrees of availability. Intuitively, this is just what membership grades denote.

It's an attractive feature of this approach that we are free to interpret the membership grades in different (not necessarily competing) ways. For example, we can interpret them as denoting distances from possible worlds: the larger an option's membership grade, the further away the nearest possible world in which it isn't available. This interpretation would fit republican or independence-based conceptions of freedom, such as those defended by Pettit (2001) and List and Valentini (2016), according to which an individual is free to ϕ if and only if there is a *robust absence* of externally imposed constraints on their ϕ-ing (i.e., the nearest possible world in which constraints are imposed is quite distant). Following my arguments in section 2, however, I propose to interpret membership grades as conditional probabilities. In other words, I interpret $m(x)$ as the chance that an agent can successfully choose x if they intend to.

Obviously, the value of $m(x)$ will depend on several factors; agents always face particular constraints (physical and mental limitations, poverty, lack of skills or education, geographic isolation, coercion, etc.) that simply rule out particular options, and as we saw earlier there are other factors which conspire to make options less accessible (whether these options require the recruitment of motivation to be chosen, how much motivation is required, how strong-willed the agent in question is, what other options are available, and how salient each option is), without ruling them out entirely. Accordingly, a set's membership function is actually an n-argument function, where each of the n arguments denotes one factor that impacts option accessibility.

Note that the value of an option's membership grade need not track very closely how effortful it is to choose that option: if you are iron-willed, and you always succeed in choosing as you resolve to or as you judge you ought, even when faced with such temptation as requires strenuous effort from you to resist, then my approach might count all your options as fully accessible. This seems right: if you can always overcome temptation when you want to, then temptation is no constraint upon your freedom. You may *prefer* to be rid of temptation, but then being exposed to temptation simply means that the level of welfare you achieve is lower than it might otherwise have been, not that you are any less capable of doing, being, or choosing anything. Likewise, if you are incredibly weak-willed, and abandon your resolutions as soon as it becomes hard to maintain them, then the membership grade of some of your options will be very low. Again, this seems right.

Observe also that while it is the problem of weakness of will that motivates my proposal, my framework is perfectly general. It imposes no constraint on the list of factors that can be included as arguments of a set's membership function, nor does it require the inclusion of any particular factor, and it allows for flexibility in the way that we can represent the impact of these factors that we do include on people's freedom. Notably, my framework permits us to include external constraints among the arguments of a set's membership function, and to represent the effect of these constraints as that of merely diminishing the accessibility of an option. Remember that because standard approaches to the measurement of freedom assume bivalence in set membership, it is impossible for a constraint upon one's freedom to *be* a constraint unless it entirely rules out certain options. But consider the following case devised by Garnett (2007, p. 438): you wake up in a room, and you find the door is locked by a padlock whose combination you don't know. Is the option to leave the room available to you? Intuitively, it's hard to say: on the one hand, all you need to do to get out is put in the right combination, but on the other hand, if all you can do is guess at possible combinations it looks like you may never get out. My account deals with this case rather elegantly: the option to get out *is* available to you, but not very accessible, because, conditional on your intending to put in the right combination (whatever it is) and get out, the chance that you will do so is vanishingly small. And naturally, if you are somehow able to narrow down the list of possible combinations, then my analysis implies that your freedom has just increased substantially, since the chance that you will get out soon, given that you intend to, has just increased.

This analysis of probabilistic external constraints is interestingly distinct from other analyses that have been proposed. Carter (1999, p. 191) also argues that a measure of freedom must incorporate probabilistic judgments in some way, but his proposal is that we make an individual's freedom increasing in their expected number of options. The expected number of options available to an individual is simply the sum of all possible options, discounted by the *unconditional probability* of there being an *externally imposed preventing condition* on the option which makes it *fully unavailable*. Importantly, this proposal retains bivalence, in that for any option x and any opportunity set A, either x is fully in A or x is not in A, it's just that there's some uncertainty over which A will in fact be yours. This proposal has substantively different implications than mine, for the simple reason that the presence or absence of a preventing condition on my ϕ−ing need not be probabilistically independent of my intention to ϕ.

Imagine that I have a sworn enemy who is a mind-reader, and who will attempt to prevent me from carrying out whatever intention he reads in me. In that case, the *conditional probability* that there will be a constraint on my ϕ-ing, given that I intend to ϕ, will be much higher than the conditional probability that there will be a constraint on my ϕ-ing, given that I don't intend to ϕ. So there being a constraint on my $\phi-$ing is not probabilistically independent of my intending to ϕ. In this situation, for any ϕ, the chance that I will do ϕ, given that I intend to, will be very low, since there is in that case a *high* chance of a preventing condition. In contrast, for any ϕ, the *unconditional probability* of there being a preventing condition on my ϕ-ing may be quite *low*, since my enemy won't bother to impose constraints on my $\phi-$ing if they don't discern in me any intention to ϕ. It's only *if* I intend to ϕ that the chance of there being a constraint on my ϕ-ing increases. Carter's proposal would imply in this case that I enjoy a good deal of freedom, almost as much, in fact, as if I had no enemy bent on thwarting my will; my proposal, however, implies that I enjoy very little freedom, in the sense that anyone with a very small number of fully available options would enjoy greater freedom than I. And this, in fact, seems right. So not only is my analysis of probabilistic external constraints different than Carter's, it has more plausible implications.[6]

Most importantly, however, my framework also allows us to refine the standard ranking rules surveyed in section 1 in ways that are sensitive to the heterogeneity in the capacity of individuals to carry out their choices. To see this, just suppose that in the case of Willy and Regina, all four of Regina's options are fully accessible, but Willy only has one fully accessible option (going out to drink) and his three work options are only partially accessible. Let $\succsim$ be defined over Z^*, the fuzzy analogue to the simple cardinality rule now tells us that for any two opportunity sets A and B belonging to Z^*, $A \succsim B$ just in case $|A| \geq |B|$.

Let A, B denote, respectively, Regina and Willy's opportunity sets. Whereas the simple cardinality rule would tell us that $A \sim B$, since Willy and Regina both have four options available, the fuzzy cardinality rule tells us that $A \succ B$, because $|A| = 4 > |B| > 1$. And if we assume that by depriving Willy from the choice to go out drinking, we increase the membership grade of his three work

[6] Matthew Kramer (2003, p. 175) also argues that the ascription to an agent of the freedom to ϕ must carry "a probabilistic qualification" indicating what the chances are that the agent will enjoy that freedom, however he nowhere states explicitly how chances are to be interpreted, or what form exactly the qualification is to have, so he is difficult to locate in this conversation. That said, since he is a proponent of pure negative freedom, like Carter, and believes that one is unfree to ϕ just in case other agents impose preventing conditions on one's ϕ-ing that make it impossible for one to ϕ, I feel it is fair to impute to him the same view as Carter regarding how the uncertainty of there being constraints on one's actions should affect the measure of one's freedom. To the extent, then, that Kramer and Carter agree with each other, my view is also in disagreement with Kramer's.

options, the fuzzy cardinality rule tells us that Willy's freedom increases provided that the change in the sum of the membership grades of the three work options is greater than 1. This rule will also rank Cassandra's opportunity set as offering strictly more freedom than Perseus's. Thus the fuzzy cardinality rule generates more intuitively correct rankings than the simple cardinality rule.

Any other ranking rule can similarly be generalized to an ranking rule over Z^*. It is easy, for instance, to incorporate a minimal desirability constraint on the formula above, by stipulatively assigning a membership grade of 0 to any option which fails to satisfy the desirability constraint. Considerations of similarity can be incorporated in similar fashion. A extremely simple proposal might run as follows: instead of counting how many (minimally desirable) options in an opportunity set, we could count how many (minimally desirable) option types there are, and identify the membership grade of each option type with the membership grade of the most accessible option token of that type.

There is much to be said in favour of my proposal, then, to represent opportunity sets as fuzzy sets. It is a natural generalization of the existing approaches to the measurement of freedom, but the great advantage of this approach is that is allows us to take account of the fact that options may be more-or-less accessible. As demonstrated, it allows us to reach more accurate measures of individual freedom.

2.5 Application

One of the main motivations for developing a measure of freedom is that the concept of freedom is central to many debates in political philosophy and public policy. In this section, I briefly explore how one might try to use my framework to generate new insights into old questions by applying it to the ethics of nudge.

Nudging consists in changing the way in which choices are presented to people so as to induce them into making choices that are better for them, or that are thought to be better for society. Crucially, nudges operate by exploiting people's reflexive choice habits, their unconscious, irrational tendencies to choose in particular ways when choices are presented to them one way rather than another (Bovens 2009). No steps are taken to remove any options from individuals, nor to impose any burdens on individuals who rationally choose the behaviors we are aiming to discourage, nor to provide them with new information that might cause them to consciously revise their behaviours (though nudging may supplement information campaigns).[7]

The no-burden requirement is a bit ambiguous, but what Thaler and Sunstein (2003), for example, have in mind is that we shouldn't make it more expensive or more time-consuming to engage in the sort of behaviour we're trying to discourage, because this would make it impossible for individuals to engage in it *and* spend this money or time on other things. In a way, then, the no-burden requirement can be interpreted as re-emphasizing the crucial point that nudges must not make any previously available option unavailable, where options are conceived here, as before, as conjunctively possible courses of action.

A paradigm case of nudging is *save more tomorrow*, which was a program proposed by Thaler and Benartzi (2003) to increase employee's contributions to their 401(k) retirement plans. Under this plan, employees were asked some time *before* receiving their next pay raise whether or not they wanted they wanted to commit this raise to their pension plan, as opposed to being asked once they received their raise. This plan tends to greatly increase savings rates: in the first company to participate, savings went from 3.5% to 11.6%. As Bovens (2009) explains, *save more tomorrow* exploits two design flaws in human psychology: first, the endowment effect. People find it harder to part with what they already have than with what they don't yet have. Second, people find it harder to resist temptation than to make provisions for the future so that they aren't tempted at all. Employees do remain, of course, free to choose either way. So, although nudging is paternalistic, to the extent that we are attempting to guide people's behaviour in ways that we judge to be best for them, Thaler and Sunstein (2003) argue that it is a benign, "libertarian" form of paternalism, because it does not reduce the available range of choices.

Thaler and Sunstein's argument has proven controversial. Hausman and Welch (2010), argue that the "libertarian" credentials of nudges are dubious, because though nudges do not remove options, they may undermine individual control over their choices and evaluations, by making individuals act in ways that reflect the social engineer's designs, not the individual's. Bovens (2009, p. 218), likewise, though he supports some nudges, is more circumspect than Thaler and Sunstein, arguing that nudging is problematic if it aims to make us choose in ways that are not in line with our actual preferences. He gives the example that society may be better off if I am nudged not to place an additional fishing boat in already over-fished waters, but that I may be worse off for being so nudged. This case raises some concerns, he believes, because we are being nudged to choose in ways that

[7] Thaler and Sunstein (2003) and Sunstein (2019) actually count providing individuals with information as nudging them, but Bovens (2009) and Hausman and Welch (2010) argue that this is a mistake.

are aberrant, out of touch with our overall judgements of what is in our interest Bovens (2009, p. 213). In contrast, he argues that nudge is much less worrisome if it brings our agency into better alignment with our preferences; notably, in cases where we are limited by ignorance, weakness of will (I will come back to this in a minute), status quo bias, or some irrational queasiness from making the choice that best fits our preferences, nudge may induce us to choose in the way that, on reflection, we judge is best.

My framework offers a rather new perspective on this debate. As I will show, it is not in general true that nudge doesn't affect the degree of individuals' freedom. In particular, it is certainly false to claim that nudging never reduces individual freedom. More surprisingly, however, my framework implies that in some cases, nudge may actually expand individual freedom. Indeed, the most attractive nudges will be those that do expand individual freedom.

Consider first Thaler and Sunstein's response to the objection that nudging is paternalistic. If the main reason that paternalism is objectionable is that it curtails the freedom of individuals against their consent (Arneson, 1980), then it won't do to point out that nudge does not restrict the range of choices available to agents: as we've seen, this is no guarantee that nudging does not restrict individual freedom. Consider, for instance, the "Don't mess with Texas" campaign, which aimed to reduce littering by saturating road signs and advertisements in radio and in print with the phrase "don't mess with Texas," often featuring popular celebrities like Stevie Ray Vaughan in their ads to drive the slogan home. As Hausman and Welch (2010, p. 134) point out, although the campaign was informational (it ran and continues to run educational campaigns to teach Texans about the harms of littering), its central slogan and messaging "attempted to create a machismo image for those who don't litter," in essence exploiting Texans' self-image to make them want to not litter. As any standard model of behaviour in the economics of identity (e.g., Akerlof and Kranton, 2000a) would predict, the effect of successfully portraying some behaviour as being prescribed by individuals' conception of their identity will be to generate a desire to engage in it.[8] This desire, of course, may run contrary to one's better judgment or to one's resolutions, and for this reason it may diminish one's freedom of choice if motivation needs to be recruited to overcome it.[9]

What this example shows is that Thaler and Sunstein's "no substantial burden"

[8] More precisely, in Akerlof and Kranton's terms, individuals will suffer a loss of identity-related utility if they behave against the behavioural prescriptions which they take to follow from having the identity they ascribe to themselves – e.g., "real Texans don't litter."

[9] Perhaps this example isn't so troubling, as the aim was only to reduce littering, but suppose instead it had been to induce Texans to mass-purchase useless and expensive consumer goods.

requirement does not suffice to guarantee that nudging does not restrict freedom. Rather, two conditions must be met: (1) that no option be ruled out, nor (2) be made less accessible. These conditions are not met in the case of the "don't mess with Texas" campaign. But if both these conditions are met, then my approach to the measurement of freedom implies that individual freedom is not restricted by nudging.[10] The charge of paternalism thus loses some of its sting, as Thaler and Sunstein claim, though in fewer cases than they hoped.

More surprisingly though, and more interestingly, my framework implies that in some cases nudging may actually *expand* individual freedom. Here again we turn to cases of weakness of will. In *save more tomorrow*, the reason individuals are more prudent if asked before rather than after receiving their raise is that at t_{after} they are exposed to a source temptation (i.e., the check in their hands), but not at t_{before}, and so they more competently manage their spending decisions at t_{before}. The nudge, in this case, consists in nothing more than *removing a source of temptation from decision-makers* by changing the time at which they must make their decision, thus making the option to save their next raise towards retirement more accessible, without in the interim making it any harder for them to spend their money in other ways or changing how they may choose to spend it. In this case, a fuzzied cardinality-based ranking implies that nudging will expand individual freedom of choice, and this should take the sting out of the charge of paternalism. Given, then, that nudges which are limited to removing sources of temptation increase individual freedom, these will be the most attractive sorts of nudges.[11]

I should note that I am not the first to argue that nudge may expand freedom in cases of weakness of will. Sunstein (2019, p. 63) also defends this claim. However, on his view, it is because nudges allow individuals to achieve outcomes they prefer that they may expand individual freedom. Plainly, this explanation commits Sunstein to a strongly value-laden measure of freedom. Being free, Sunstein tells us, is (in part) a matter of actually achieving preferred outcomes. What is distinctive about my claim is precisely that it is couched in a value-neutral account

[10] In some cases, nudging might decrease the accessibility of some options and yet also increase the availability of others; for instance, by making unhealthy food choices in a cafeteria less salient, I may reduce the force of temptation they exert on dieters, but obviously this will increase the salience of healthy food items, and if people have an irrational tendency to pick the most salient options, this could decrease the accessibility of the unhealthy options. In this case, the overall effect on freedom could be a wash, though strictly speaking (2) is not satisfied. We can perhaps reformulate it as the condition that no option be made less accessible without compensating increases in the accessibility of other options.

[11] A potentially more chilling implication of my approach is that subliminal messaging could also be freedom-enhancing, if the subliminal messages were designed to manufacture desires that helped us overcome weakness of will. Perhaps implanting in people a desire to be resolute would make them better at resisting temptation.

of what it means to be more-or-less free to pursue a given course of action. And this, I submit, makes my conclusion more surprising than Sunstein's: it is not news that exploiting certain features of people's psychology may make them better off – that much was already clear from examples like *save more tomorrow* – but it is somewhat surprising that this could make them more capable agents.

In sum, by looking at the ethics of nudge through the lens of my framework, we are better able to judge when nudging is compatible with freedom. More could be said by looking at additional cases, but I hope this brief discussion has shown why my framework might be of interest in particular cases.

Summary

Cases of weakness of will reveal that measuring individual freedom requires us to take account of the heterogeneity in the accessibility of options. Because the assumption of bivalence is baked into extant measures of freedom, they lack the machinery with which to take this heterogeneity seriously. Thus, otherwise attractive ranking principles generate bizarre rankings in cases where weakness of will gets involved.

My proposal resolves this problem. Representing opportunity sets as fuzzy sets allows us to model the fact that option accessibility admits of degrees. And in redefining the domain of the "offers at least as much freedom as" relation to the set of all "fuzzied" opportunity sets, we can generalize the standard models so as to both preserve what is attractive about our preferred ranking rules, and also generate the intuitively correct rankings in the problem cases involving weakness of will. This framework has the advantage of being at once more precise in its treatment how weakness of will constrains human freedom than other accounts, while remaining perfectly general, providing an attractive analysis of constraints that applies equally well to internal constraints as to external ones. Finally, this framework's appeal also lies in its potential for new insights into other problems, as illustrated in the case of the ethics of nudge, and as we will see in the next chapter, where I apply this framework to define measures of freedom that are sensitive to how robust one's opportunities are to counterfactual interventions.

Chapter 3

Fuzzying Modal Conceptions of Freedom

In *The Idea of Justice*, Amartya Sen (2009, p. 306) draws our attention to two different contrasts in relation to individual freedom. First, there is the contrast between being capable of doing something, and being incapable of doing it. This is the contrast which Sen's capability approach aims to capture, and he illustrates it by comparing the situation of a disabled person, Alice, who needs help to perform various tasks but has no one to help her, say, because local gangs scare off aid workers from coming to the neighbourhood (case 1), with that of Alice's situation when there are no gangs and some local charity or social security program organizes caregivers to help her do what she could not previously do (case 2). Second, there is the contrast between being independent of the will of others, meaning that others *could not* hinder one's purposes if they tried, and being dependent on their will, in which case they could. This is the contrast which the republican model of freedom aims to capture (Pettit, 2012), and Sen illustrates it by comparing Alice's situation in the previous two cases with Alice's situation in a case where she employs well-remunerated servants who obey her commands (because *they* are dependent on *her*) and allow her to do what she wants to (case 3).

Importantly, according to Sen, the capability approach does not register any difference in the second and third cases – in both cases Alice can *do* what she previously could not – but neither does the republican model register any difference in the first and second – in neither case is Alice independent, since she can be hindered by the will of others (and in fact is, in the first case). Since both models of freedom generate inconsistent assessments of who has more freedom than whom, much debate has taken place over which offers the best conception of freedom. Sen (2009, pp. 307-308), however, invites us to see both approaches as complimentary rather than competing. Both approaches capture one contrast that is central to how we think about freedom, and are blind to another. Freedom is an inescapably plural concept, and it is not desirable to reduce it to one unitary interpretation.

Sen's insight invites the following inference: if it is undesirable to impoverish our concept of freedom by limiting ourselves to a single interpretation, it should likewise be undesirable to impoverish the measure of freedom by insisting on a single measure. I argued in Chapter 2 that opportunity sets should be represented as so-called "fuzzy" sets, and I hinted that the membership grades which define set membership in fuzzy set theory might admit of two different interpretations, one probabilistic, and another modal. Having treated the probabilistic interpretation in Chapter 2, in this Chapter I explore the modal interpretation, and I argue that it provides an attractive measure of republican freedom – indeed, it provides an attractive way of measuring freedom on *any* view of freedom which takes modal considerations seriously, and this is a class of views which also includes freedom-as-independence, and even some interpretations of the liberal view of freedom.

My plan is as follows. In section 1, I lay out the republican model of freedom in greater detail, contrasting it briefly with the liberal and independence models, and show how modal considerations matter on all three of these views. In section 2, I leverage a Lewisian account of counterfactual semantics to construct a (partially) cardinal measure of distance over possible worlds – this is, to my knowledge, the first such measure ever proposed – and I show that this measure allows us to construct a further measure of how robustly available an option is. This second measure gives a meaningful modal interpretation to membership grades, and this allows us to define measures of freedom that helpfully characterize the commitments of modality-sensitive conceptions of freedom. This is, to my knowledge, the first modality-sensitive measure of freedom, and therefore the first measure to capture republican concerns. In section 3, finally, I reflect on the contrast between the measures developed here and those developed in Chapter 2, and concludes that Sen is right to insist on the plurality of freedom.

3.1 Republicanism and Independentism

Pettit (1997) introduces the republican conception of freedom as the view that freedom consists in being *undominated.* Someone dominates another, according to Pettit (1997, p. 52), when they have the capacity to interfere arbitrarily in their choices. Accordingly, one is free to the extent that others cannot arbitrarily interfere in our choices (through physical coercion, threats of violence, blackmail, manipulation, etc.). This conception of freedom has three essential features.

First, it is negative, in the sense that it considers that one is free to ϕ just in case there is an absence of constraints imposed by others on one's ϕ-ing. Second, the conception of constraints at issue here is a *modal* one. For Pettit, it is not enough that no one *actually* interfere arbitrarily with our choices – they must not even have the possibility of doing so. The absence of constraints on our ϕ-ing must be *robust.* List and Valentini (2016, p. 1048) helpfully clarify this requirement:

> "[F]or an agent to be free to do [ϕ], the following must be true: for each possible world w within a sufficiently large class of worlds accessible from the actual world, there exists some possible world w', accessible from w, in which the agent does [ϕ]. It is important to emphasize the double modality here: the possibility of doing [ϕ] must exist not only in the actual world, but for each world w within a relevant class of worlds accessible from the actual world."

The classic example of a person who is not robustly free in the sense required here is that of a slave with a kind, non-interfering master. The master, we can suppose, lets the slave do what they will, does not force them to do any work, and even gives them an allowance – think Dr. King Schultz from *Django Unchained,* who buys Django and keeps him as a slave just long enough to track down the Brittle Brothers, but who pays Django for his service, does not force him to do anything against his will (both want to kill the Brittle brothers), and frees him afterwards, as promised. In this context, the slave is not being interfered with, nor is he even *likely* to be interfered with, but they clearly remain fettered in intolerable bondage, since it would be the easiest thing in the world for the master to assert their power and interfere with them at will. By merely retaining this power of absolute control, slave owners carve unjustifiable inroads into the freedom of their victims.

It is worth noting that for Pettit, the freedom to ϕ presupposes the capacity to ϕ, not merely the absence of constraints, but it is only the absence of constraints on one's ϕ-ing which he requires to be robust, not our capacity to ϕ. Thus, supposing you can in fact ϕ, you count as *free* to ϕ, on Pettit's view, just in case, in none of the worlds within the relevant modal base, does anyone impose *rob* you of this capacity. Note that this is consistent with you not having the capacity in each of these worlds. There may be nearby worlds in which you break a leg and thereby become unable to scale a mountain, but that in itself does not impair your freedom in the actual world. List and Valentini don't express any view about this asymmetry, so I shan't impute it to them.

Third and final, Pettit's conception of freedom is *moralized.* Only the *arbitrary* power to impose restrictions on others is injurious to their freedom. If, although an official has the power to interfere with the choices of others, their use of this power is constrained by procedures designed to take account of the interests of the target of the restrictions, then the restrictions they impose do not count as freedom-restricting (Pettit, 1997, p. 55). This is the crucial point of contrast between

Pettit's republican model of freedom and List and Valentini's independence model of freedom, which is non-moralized, and holds more simply that one is free to ϕ just in case there is a robust absence of constraints on our ϕ-ing.

List and Valentini (2016, p. 1049) argue that we should not moralize our notion of constraint, because it is an important *desideratum* of a conception of freedom that it pick out those constraints (either actual or possible) which stand in need of justification. Authority and power are not self-justifying, even if they are themselves subject to procedural constraints and exercised with an eye to the interest of those who are subject to them. The independence model of freedom recognizes this, and marks all forms of power and authority as sources of unfreedom. In a way, List and Valentini suggest that freedom works as a kind of conceptual alarm-system: observing that, say, a policy or an institution infringes upon individual freedom (or would infringe upon it if implemented) should serve to give us pause, force us to scrutinize the reasons that speak against or in favour of said policy or institution, etc. The republican conception of freedom, by refusing to count justified constraints as constraints at all, builds too much into its analysis of freedom from the get-go to serve this important alarm-system function.

The heavy emphasis on robustness that is common to both republican and independence conceptions is supposed to be what sets them apart from the more classical liberal conception of freedom defended, most prominently, by Steiner, Carter, and Kramer. As we've seen, the liberal view holds that you're free to ϕ just in case there are no actual constraints on your ϕ-ing; if someone can easily impose constraints on your ϕ-ing, but doesn't, this isn't supposed to constrain your freedom. And this is borne out by Carter's measure of freedom, which is an increasing function of the ratio of the expected number of acts you can perform to the expected sum of the number of acts you can perform and the number of acts you are prevented from performing. This measure trivially implies that a slave with a kind master enjoys as much freedom as a free person, provided the master is unlikely to interfere; this shows why, according to List and Valentini, the liberal conception of freedom is inadequate.

And yet, at the same time, Steiner, Carter, and Kramer all evince a concern for so-called "subjunctive" constraints. That is, they all concur that if there are currently no preventing constraints on your ϕ-ing, but there would be preventing constraints on your ϕ-ing if you *attempted* to ϕ, then you aren't free to ϕ. But note that this concern for counterfactually dependent constraints isn't captured by their measures of freedom. Carter's measure of freedom depends on expectations, not conditional expectations: options are discounted by the *un*conditional probability of being unavailable, not their conditional probability of being unavailable, given that you intend to exercise them. And as we saw in the last chapter, the *un*conditional probability of there being preventing constraints on my ϕ-ing might be very low, even if the conditional probability that I would be unable to ϕ *were I* to attempt to ϕ is very high. In these situations, Carter's measure implies I enjoy considerable freedom to ϕ, but this flies in the face of the concern for counterfactually dependent constraints.

One way of capturing this concern is in the way I did last chapter, where we essentially "discount" your freedom to ϕ by the *conditional* probability that you will fail to ϕ, given that you intend to ϕ. But another way of capturing the concern for subjunctive constraints is through a weak modal safety condition: you are free to ϕ if and only if you can ϕ (i.e., the option to ϕ is available) and, moreover, this isn't simply by complete chance, e.g., by mere dint of the fact you don't happen to want to ϕ. Thus being free to ϕ requires that ϕ-ing remains an available option in nearby worlds. This would amount to accepting List and Valentini's definition of what it is to be free to ϕ, but making the modal requirement very weak, thus preserving a contrast with republican and independentist views, which typically impose very strong modal safety constraints. This is a non-standard interpretation of the liberal conception, but one which captures an avowed concern of liberals that is not captured by their own measures.

In sum, modal considerations are salient on a broad range of major views on the nature of freedom. Therefore, a measure of freedom that is sensitive to modality seems necessary if one wishes to carry out a sensible comparison of who enjoys more freedom than whom.

3.2 Modality and freedom

3.2.1 The distance between possible worlds

Before we can define a modality-sensitive measure of freedom, we first need a measure of robustness. List and Valentini explain what it is for an option to be robustly available, *simpliciter*, or fully robustly available: it is for it to be available in every world within some sufficient distance d from the actual world. But presumably, if an option isn't available in every world that is within d of the actual world, it will nonetheless be *more* robustly available the further away from the actual world is the nearest possible world in which it isn't available. To validate the claim that some option is more robustly available than another, we need some way of assessing the distance of possible worlds from the actual world.

In many cases, these assessments are intuitive enough. To come back to the issue of the slave and the non-interfering master, it's clear that although the master may allow the slave to ϕ, the world would not have to be very different for the slave to be unable to ϕ – the master could simply change their mind – and this is what allows us to say that the slave is not robustly free to ϕ. And again, it's worth remembering that while the master might not be *likely* to change his mind, it would be *easy* for him, and this is why the slave is not robustly free to ϕ: it's too *easy* for someone to prevent them from ϕ-ing; their ability to ϕ is not resilient to changes in the world. In contrast, if we consider a citizen of a robust liberal democracy currently engaged in a protest movement, it's clear that the world would have had to be quite different for this citizen to have been unable to protest the government. So the world in which the free citizen is interfered with is more distant from the actual world than the world in which the slave is interfered with. In fact, the nearest world in which the free citizen is censored seems *much* more distant from the actual world than the world in which the slave is interfered with. Our intuitions in this case are *so* sharp that they suggest a cardinal notion of "distance between possible worlds." In other cases, it's less clear which is the nearest possible world: is the world in which Bernie Sanders won the 2015 democratic primary and then beat Donald Trump closer or farther away from the actual world than the world in which Bernie won the primary but lost the general election?

The seminal treatment on assessing the distance of a world from the actual world comes from Lewis's (1979) account of similarity between possible worlds. Following Stalnaker (1968), Lewis invokes the concept of similarity between worlds in order to validate counterfactual statements. According to Lewis, counterfactual propositions of the form " if A were the case, then B would be the case" (denoted $A\Box \longrightarrow B$) are true if and only if there is some possible world in which A and B are both the case which is more similar to the actual world than any world in which A was the case but B failed to be the case. Thus, according to Lewis, the counterfactual "if Nixon had pressed the button, then there would have been a nuclear war" is true because there is a world in which Nixon pressed the button and there was nuclear war which is more similar to the actual world than any world in which Nixon pressed the button but no nuclear war ensued (Lewis 1979: 465).

This analysis is intuitive: when we reason counterfactually, we ask ourselves what we would expect to have happened, given our knowledge of the laws and other background conditions, had some counterfactual event occurred. In effect, we're asking ourselves what the world would look like if we changed the least about it that was necessary for it to have been the case that the counterfactual event occurred. And indeed, if you think that C would have occurred had A been the case, but you learn that in fact B would have occurred instead, what you've learned is that the world is different from how you thought it was.[1]

What determines the similarity order? Lewis's answer to this question is much more controversial than his truth conditions (see e.g., Slote, 1978; Edgington, 1978; Wasserman, 2006; Kment, 2006), and boils down to the degree to which two worlds agree on laws and facts.[2] Thankfully, we don't need to get into the weeds of *what* determines the similarity ordering, since most of Lewis's critics don't disagree that some considerations *do* determine a (partial) similarity ordering. What *is* important for our purposes is that if Lewis's analysis of counterfactuals is on the money, then we can construct a (partially) cardinal distance measure of possible worlds from the actual world representing their degrees of dissimilarity, meaning that not only can we rank possible worlds according to how distant they are from the actual world, but we can also rank differences in distance from the actual world. The full procedure is described and illustrated in appendix A; I only summarize the main results here.

The basic idea is that since true counterfactuals are validated by facts about which worlds are more similar than which to which others, we can infer, from the (infinitely large) body of true counterfactuals, a preorder $\trianglerighteq$ on the set $\mathcal{W}$ of possible worlds, ranking possible worlds by dissimilarity from the actual world w_a, and a preorder $\trianglerighteq_D$ on $\mathcal{W} \times \mathcal{W}$, ranking pairs of worlds by their difference in dissimilarity. The existence of this quaternary relation follows from the semantics of embedded counterfactuals (see appendix), and the demonstration of its existence is significant, metaphysically, because it expresses *cardinal* relations of world similarity. This should come as a surprise to metaphysicists, who have generally held that world-distance is (at best) ordinal (Morreau, 2004). We can therefore define a

[1] A study by Stanley, Stewart, and Brigard (2017) attests to the intuitive character of this analysis, strongly suggesting that judgements of comparative world-similarity play a large role in judgements of plausibility in counterfactual simulations. I should note that Stalnaker's analysis of counterfactuals, which also invokes the concept of world-similarity, gives counterfactuals different truth-conditions, and some philosophers prefer Stalnaker's, but I am focusing on Lewis here because his analysis has been more influential. This is innocent enough: the argument I run with Lewis's truth-conditions could equally be run with Stalnaker's.

[2] In the Nixon example, although we haven't had a nuclear war – and in this respect the actual world is more similar to the world in which Nixon's button-pressing didn't cause a nuclear war – for Nixon's pressing the button to have failed to cause nuclear war, widespread violations of natural law would have had to occur (the missiles were magically filled with confetti instead of nuclear payloads, the launch system suffered a mysterious malfunction...). Moreover, Lewis goes on to explain, although the series of events that would have followed a nuclear strike would be quite different from the series of events that actually took place, the series of events that would have taken place had Nixon pressed the button but failed to launch a war would also have diverged, if only slightly, from the actual series of events (he would have left digital prints on the button, the click of the button would have been preserved on tape, his thought process would not have been the same, etc.), barring further, gratuitous violations of natural law.

distance function $d : \mathcal{W} \to \mathbb{R}$ representing both relations, and, if we set $d(w_a) = 0$, then d is unique up to positive affine transformations. Our function d has as many arguments as there are dimensions of similarity, so if there are n dimension, $\mathcal{W}$ is isomorphic to the n-dimensional Euclidean space E^n.

I should confess that the construction described in the appendix assumes that $\trianglerighteq$ and $\trianglerighteq_D$ are both complete, that is, that there is a unique, fully determinate similarity order to be inferred from the body of true counterfactuals. This would not sit well with most metaphysicists. World-similarity is notoriously vague, due to the context-dependence of counterfactual semantics: as Lewis points out, there are ways of filling in the details of counterfactuals that might make it true that if Caesar had led the U.S. side in the Korean war, he would have used nukes, but other ways of filling in the details that make it false. Lewis's analysis thus implies different rankings of similarity and differences in similarity between possible worlds under different resolutions of the underlying vagueness. Each of these possible rankings would be complete, but on pain of arbitrariness in our choice of similarity relations, and therefore in our choice of distance function, a possible world w can only be unambiguously more dissimilar from the actual world than some possible world w' if every admissible dissimilarity relation agrees on it. Therefore, world-distance can only be measured by a *family* $D = \{d_1, d_2, ...\}$ of distance functions, each unique up to positive affine transformations.[3]

3.2.2 Another fuzzy proposal

We can now turn to defining a modality-sensitive measure of freedom. Following List and Valentini's definition, let's assume that if an option is available in every world within a sufficiently large neighbourhood of the actual world, then it is maximally robustly available. More precisely, let N be a convex subset of $\mathcal{W}$ with a least upper bound, meaning that there is a world w such that any world further out from the actual world than w does not belong to N, and assume that if some world w belongs to N, then any other possible world w' which is not further away from the actual world than w also belongs to N. We then say that if an option

[3] This holds even on an epistemicist theory of vagueness (Williamson, 1992). According to epistemicism, all predicates, including modal predicates like "would've used nukes," actually do draw sharp boundaries, and their apparent vagueness is due only to our own (unimprovable) ignorance of where exactly their extension lies. Epistemicism implies that actually counterfactuals aren't *really* vague, and therefore there *are* unique relations $\trianglerighteq$ and $\trianglerighteq_D$. Nonetheless, under epistemicism, we can't know what the unique relations are, we only know that they are members of some more-or-less large family of possible relations, each of which is consistent with what we do know about counterfactual facts. Or in other words, we only *know* that w_1 is, say, at least three times further from the actual world than w_2 if every admissible distance function agrees with this assessment.

is available to an agent in every world within N, then the option is maximally robustly available to the individual. If, in contrast, the option is not available in the actual world, then it is minimally robustly available. Finally, if an option is available in the actual world, but not in every world in N, let us assume that the degree of robustness with which it is available is linear in the distance from the actual world of the nearest possible world in which it isn't available any longer.

This last requirement is not obligatory: an alternative would require that robustness be linear in the measure of the largest connected set $S \subseteq N$ that only includes the actual world and worlds in which the option is available. In other words, in the representation of $\mathcal{W}$ as a euclidean space, S is the largest geometric shape you can draw around the actual world that includes only worlds in which the option is available, and the measure of this set is its n-dimensional area.

I am unsure which of these two alternative formulations is best; the first fits the intuition that an option is more robustly available the larger the minimal change you would have to make to the world to make it unavailable – i.e., the *harder* it is to render it unavailable – while the second fits the intuition that an option is more robustly available the more ways there are of changing the world without making the option unavailable – i.e., the more *resilient* your option is to perturbations in the way the world is. Both formulations capture something important about our intuitions regarding robustness, but they disagree.

The second interpretation has one significant advantage, which is that it squarely avoids a problem raised by Carter and Shnayderman (2019) for modality-sensitive measures of freedom: there is always, say Carter and Shnayderman, a very nearby but highly unlikely world in which you are killed and are thereby robbed of all freedom, thus if we measure robustness by the distance from the actual world of the nearest possible world where some option isn't available, then it turns out all our options are equally unrobustly available, which is a pathological result for a modally-sensitive measure of freedom. On the second interpretation, this problem doesn't arise, because even if there is a very nearby world in which neither x nor y are available, the largest connected set of worlds (including the actual world) in which x is available might still be very large, and (much) larger than the largest connected set of worlds (including the actual world) in which y is available.

Of course, Carter's worry only applies with full force to List and Valentini's unmoralized conception of freedom; one could also deal with the objection by defaulting back to a republican view and saying that it's not the nearest possible world that matters, but the nearest possible world where someone arbitrarily/uncontrollably/unconstitutionally robs you of some option. In this case, the first interpretation offers a perfectly workable measure of robustness for republican applications. So perhaps the first interpretation works best for republicans, and the second for independentists, who cannot draw on the arbitrariness criterion, and who anyway were the target of Carter's objection. For simplicitly, I will assume the first interpretation, but nothing in what follows hangs on this choice.

From here, as in 2, I define an opportunity set as a fuzzy set; that is, as a pair (X, m) where X is the set of all possible options, and m denotes a membership function that assigns a membership grade of between 0 and 1 to every element in X, in this case denoting the element in question's degree of robustness. Thus a membership grade of 0 indicates that the option is not available in the actual world, a membership grade of 1 indicates that the option is available in every possible world belonging to N, and a membership grade of $n : 0 < n < 1$ indicates that the ratio of the distance $d(w)$ of the nearest possible world w in which the option isn't available to the least upper bound of N is n. Since distance between possible worlds is cardinally measurable (along any particular membership function) and the "zero-one" I've used to calibrate the measure of robustness preserves interval differences, differences in membership grades respect differences in distances of possible worlds from the actual world. Z is the (classic) set of all opportunity sets.

Because there are many admissible measures of distance between possible worlds, there are accordingly, for any set $A \in Z$, many admissible membership functions. Let $M_A = \{\mu_{A1}, \mu_{A2}, ...\}$ be the family of admissible membership functions for A. With respect to any membership function $\mu_{Ai} \in M_A$, the cardinality of A is $\sum_{x \in X} \mu_{Ai}(x)$. Finally, let $\succsim$ denote a binary, transitive, reflexive, and nonsymmetric relation on Z, interpreted as the "offers at least as much freedom as relation," with $\succ$ and $\sim$ denoting its asymmetric and symmetric parts, respectively.

It may be asked why robustness is here defined with respect to some bounded space N rather than the whole space $\mathcal{W}$ of possible worlds. Quite simply, it is because robustness is bounded, but distances between worlds are not: there is a limit to how robustly available an option can be – a limit that is reached if it is available in every possible world – but there is no limit to how far from the actual world a possible world can be. And unfortunately, if there is no limit to how far a possible world can be, one cannot simultaneously require that the robustness with which an option is available increase linearly in the distance from the actual world of the nearest possible world in which it is not available *and* that there be a limit to how robustly an option can be available. But we need both assumptions to define membership grades that cardinally represent robustness. We could consistently require that an option be maximally available if and only if it is available in every world in $\mathcal{W}$, and that robustness increase in the distance of the nearest possible world in which options aren't available, but then membership grades can only ordinally represent robustness of availability. Note that in this case $|A| \geq |B|$ if and only if $\mu_{Ai}(x) \geq \mu_{Ai}$ for all $x \in X$ and all $\mu_{Ai} \in M_A, \mu_{Bi} \in M_B$. So we're left picking our regret: either sacrifice generality for precision of measurement, or insist on maximum generality and settle for ordinal measurements of robustness.

I have opted for the first route. This is not entirely for reasons of convenience. There may be independent reasons for wanting to restrict the "size" of the neighbourhood in which robustness is assessed. List and Valentini already emphasize in their definition that the possible worlds with respect to which we assess robustness have to be "accessible from the actual world." The thought seems to be that it only matters that an option isn't available to you in some other possible world if that world could be or could have been brought about by someone's voluntary actions or by natural chance, and in a way that doesn't contravene natural law. It doesn't matter to your assessment of your freedom, for example, that the option not to go to church isn't available to you in a possible world where a malevolent deity forces you to go to church. Robustness of availability is only an interesting concept insofar as it enables us to compare how well *feasible* alternative social arrangements protect individual freedom, so it is quite natural to restrict the space of possible worlds with respect to which robustness is assessed to those worlds which are "accessible" from ours. And it is fairly natural to think that there *is* a limit to how different from the actual world a possible world could be while remaining accessible from it.

With these preliminaries in hand, we can at last define modality-sensitive measures of freedom. For illustrative purposes, I'll allow myself two simplifying assumptions, to be suspended shortly: first, that individual freedom is increasing in the number of options that are actually available. Second, that individual freedom is linearly increasing in the robustness with which one's options are available – so, in particular, I'm assuming that an increase of r in one option's membership grade is "compensated for" by a decrease of r in another option's membership grade. In other words, I'm assuming that the amount of freedom offered by opportunity sets is increasing in their cardinality. Of course, we don't want the ranking that $\succsim$ induces over Z to depend on arbitrary choices of membership functions for each opportunity set, so I will further assume that A outranks B if and only if A outranks B under any admissible choice of membership function for either A or B. The modality-sensitive analogue to Pattanaik and Xu's (1990) simple cardinality rule thus states that for all $A, B \in Z$, and for all $\mu_A \in M_A, \mu_B \in M_B$:

$$A \succsim B \iff \sum_{x \in X} \mu_A(x) \geq \sum_{x \in X} \mu_B(x).$$

There are, naturally, many ways of refining this rule. If we drop my first simplifying assumption, we can embed a concern for the desirability or the diversity of options in the specification of $\succsim$. More interesting for the present context though, is that if we drop my second simplifying assumption, we can require that the degree of freedom offered by an opportunity set be strictly increasing and *concave* over the membership grade of its elements, as opposed to linearly increasing. Thus, increases in the robustness with which options are available would matter more the less robustly available the options were to begin with.

One could go on, but the basic idea is, I hope, clear enough: under any plausible conception of freedom that is sensitive to modal considerations, individual freedom will depend (in some way) on how many options are available to the individual, and on how robustly these options are available. And representing opportunity sets as fuzzy sets offers attractive ways of modelling this dependence. There are very important and instructive contrasts, however, between the robustness-based definition I've offered here of membership grades, and the probabilistic definition I offered in 2 – contrasts which validate Sen's assertion that freedom is an inescapably plural concept. So it is worth discussing some of these contrasts, and this is the task to which I now turn.

3.3 Modal or probabilistic fuzziness?

The most obvious advantage of defining membership grades modally, as representing the distance of the nearest possible world in which some course of action isn't available to the individual, rather than probabilistically, as representing the chance that the individual would perform the course of action in question *if* they intended to, is that the modal definition captures the sorts of intuitions motivating the independence conception of freedom, whereas the probabilistic definition cannot. The chance that a slave with a kind, non-interfering master would ϕ, conditional on the slave's intending to ϕ, need be no different from the chance that a free person would ϕ, conditional on their intending to ϕ. Thus a probabilistic definition of membership grades is unable to register the important contrast between the slave's situation and the free person's situation.

However, under a modal definition, since the nearest possible world in which it is impossible for the slave to ϕ is much nearer to the actual world than the nearest possible world in which it is impossible for the free person to ϕ, the membership grade of the slave's option to ϕ will be much closer to 0 than the membership grade of the free person's option to ϕ. If this holds generally across the options that are available to each, then the fuzzy cardinality of the slave's opportunity set is well inferior to the cardinality of the free person's opportunity set, thus the simple modal cardinality rule would trivially rate the free person as freer than the slave.

Likewise, to come back to Sen's original example, the probabilistic definition may indeed fail to capture any interesting contrast between Alice's situation when she is dependent on the goodwill of others to do what she wants, and Alice's situation where she employs well-paid helpers, and has the means to employ other helpers should hers decide to leave. Provided the people Alice is dependent on really are committed to helping her and likely won't drop her on a whim, the probability that Alice would successfully carry out whatever courses of action she might intend need not vary across these situations. The modal definition, however, does register the important contrast between Alice's dependence in the first case, and her independence in the second: if Alice can only ϕ with the aid of helpers, then much more would have to go wrong with the world for Alice to be unable to ϕ even though she has the means to afford help (market mechanisms would have to fail in an impressive way, or some elaborate conspiracy would have to be at play) than would have to go wrong with the world for Alice to be unable to ϕ when she is dependent on charity for help (however unlikely it is, gangs reappearing or a whim on the part of the helpers is all it would take).

Perhaps more surprisingly, some measures of freedom-as-independence – for example the simple cardinality rule – also succeed in capturing the contrast to which Sen alleged that modal conceptions of freedom are blind. An option that is available to you in the actual world is more robustly available to you than an option that is not available to you in the actual world, for the simple reason that if x is actually available to you, then the nearest possible world in which it isn't available must be *at some distance* from the actual world. Accordingly, under any measure of freedom-as-independence, the membership grades of options that are in fact available will be greater than the membership grades of options that are not at all available. Thus, such measures of freedom will register a difference between Alice's degree of freedom in the case where she has no one to help her, and Alice's degree of freedom in the case where she is dependent on the help of others. The probabilistic definition also registers a difference here, of course, but it is a uniquely attractive feature of the modal definition that it recovers our intuitive sense that Alice's freedom expands continually as we move through all three of Sen's cases.

The attractions are not all one-sided, however. A major motivation in the previous chapter for going probabilistic on the measure of freedom was that doing so allows us to represent the fact that available options may admit of degrees in accessibility. If you and I can both ϕ, but, conditional on my intending to ϕ, I am likelier to ϕ than you are, conditional on your intending to ϕ, then, I argued, the option to ϕ is less accessible to you than to me, and you must accordingly be judged less free than I am to ϕ. The probabilistic definition of membership grades guarantees this judgement. But, obviously, if membership grades are interpreted modally, then the contrast between your situation and mine washes out: the nearest possible world in which some option isn't available to me may be as distant from the actual world as the nearest possible world in which that option also isn't available to you. The trouble is that for every possible world in which the option is available to the both of us (and notably in the actual world), it is less accessible to you – but this fact cannot be represented here since no information about probabilities is embedded in the measure of robustness. Under the modal definition, therefore, the membership grade of my option to ϕ is the same as the membership grade of your option to ϕ.

A second and deeper problem for the modal definition is that it violates INS, the indifference between no-choice situations, which, recall, requires that any two singleton opportunity sets offer equally little freedom. Intuitively, if you and I are both free to do only one thing, if we both have *no choice* about what to do and are simply *forced* to do the only thing we can do, we must enjoy equally little freedom. Under the fuzzy set-theoretic representation of opportunity sets, this condition requires that if only one element in the universe X receives a non-null membership grade, then we are both equally (un)free. The probabilistic definition of membership grades satisfies this constraint, because if x is the only option that receives a non-null membership grade, its membership grade is necessarily 1: if x is the only thing you can do, then the probability you will do x is necessarily 1 (and *a fortiori* the probability you will do x conditional on your intending to do x is also 1). However, on the modal definition, the membership grade of x is not necessarily 1, or any constant value: even if x is the only thing you can actually do, there may be possible worlds where you cannot do x (but can do other things instead) and this world may be more distant from the actual world than the nearest world in which I cannot do the only thing which I can currently do. Carter and Kramer's measure of freedom violates INS as well, so this cannot be taken as a knock-down objection against a modality-sensitive measure of freedom, but it is a serious cost to pay.

Finally, there is the more practical difficulty that rules for ranking opportunity sets which, under a probabilistic definition, induced a complete order over Z, are not guaranteed to induce a complete ordering over Z under a modal definition.

Under the probabilistic definition of membership grades, there is only one admissible choice of membership function for any opportunity set, for the simple reason that chances are sharp: the probability that I will ϕ, conditional on my intending to ϕ, has a definite value. Thus, in ranking opportunity sets by their cardinality (or by some other rule suitably sensitized to diversity or quality considerations), $\succsim$ orders Z by the real numbers, which is a complete order. But since there are multiple admissible choices of membership grades for any opportunity set under a modal definition, and we must require that $\succsim$ rank A over B just in case A outranks B under every admissible choice of membership function, a cardinality-based ranking rule will not induce a complete order over Z if there is even a single pair of sets $A, B \in Z$ such that $|A| > |B|$ under one choice of μ_A, μ_B and $|B| > |B|$ under a different choice of μ_A, μ_B. An unacknowledged cost, then, of sensitizing our conception of freedom to modal considerations, is that our measure of freedom loses precision. This problem becomes particularly acute if we interpret robustness ordinally, for then one can only say that one opportunity set offers more freedom than another if, with respect to every admissible membership grade, it strictly dominates the other, and this is bound to be an extremely rare occurrence.

Summary

Lewis's analysis of counterfactuals and the framework of fuzzy set theory permit important breakthroughs in the measurement of freedom. Lewis's framework makes talk of the robustness with which options are available well-founded, as well as epistemically tractable. Fuzzy set theory offers us a natural language in which to express the commitments of modality-sensitive conceptions of freedom, by allowing us to define rankings over sets of options that are appropriately sensitive to how many options are available to individuals and how robustly available these are. Seeing as no one has ever before proposed either a cardinal measure of distance on the set of possible worlds or a measure of republican freedom, I take these to be the chapter's main contribution.

As for the question of whether the independence conception or the capability conception of freedom is the correct conception of freedom, it seems Sen still gets the last laugh. Although modality-sensitive measures of freedom turn out to be more powerful than he suggested, being capable of capturing all relevant changes in A's situation throughout Sen's three examples, they remain blind to another consideration which is highly salient in comparative assessments of freedom, namely, how *capable* people are of doing the things they can do. Meanwhile, alternate measures of freedom which *are* sensitive to this consideration are in turn blind to the modal dimension of assessment of individual freedom. And no single measure will do on its own, since each is needed to correct the other's myopia. Freedom is, indeed, inescapably plural.

Chapter 4

Quality Freedom

In the previous two chapters, we have been exclusively focused on the question of how to measure freedom. So far we've seen that the extent of a person's freedom seems to depend on how many options they have, how accessible they are, and how robust they are. This chapter aims to conclude our exploration of what determines a person's freedom, by considering how the quality and the diversity of our options bear on the extent of our freedom. By the end, I hope to arrive at a tentative final statement of what, I think, makes a person more-or-less free. However, this chapter also marks an inflection point in the narrative of our exploration, as in the course of discussing the role of option quality and diversity on the measurement of freedom, we will be brought to consider what makes freedom *valuable*, in the first place. And this will become our new topic of concern for the next two chapters.

It is a vexed subject for both philosophers and economists as to whether the quality of one's options affects how much freedom of choice they offer. Numerous philosophers in the liberal tradition (e.g., Carter, 1999; Kramer, 2003; Hurka, 2011; Bader, 2018) exclude quality considerations from their measures of freedom. These authors are often motivated by the conviction that restrictions on individual freedom should be justifiable to those individuals in terms that presuppose no contentious moral judgements. In contrast, others have argued that one is freer when the options one has are better, in the sense of being preferred by the individual, or reasonably preferable by someone similarly situated (e.g., Taylor, 1985; Sen, 1990c; Pattanaik and Xu, 1998). These authors are guided by the thought that a person is freer the more capable she is of living the sort of life that someone might reasonably want to live. If we compare people's freedoms simply by counting how many options people have, we will be unable to recover important intuitive data; for example, everything else being equal, countries which protect *important* freedoms protect freedom *more* than countries which do not. There is something to both approaches, but they don't sit well together. My aim is to develop a view of how to measure freedom and of why freedom matters, which gives both intuitions their due. As a first step, I present a new way of taking account of the quality of options when measuring freedom, based on the capacity of those options to provide individuals with the standing to be held responsible, rather than on anyone's preferences. This measure, I argue, improves in key respects on existing quality-sensitive measures of freedom.

Quality-based accounts of the measure of freedom are of two different kinds. One view, associated with Sen (1990c), Taylor (1985), and Suppes (1987), is that the more intensely an individual might reasonably desire some option, or the better off they would be if they chose it, the more this option contributes to their freedom. An obvious problem with this view is that if two individuals each have only a single option, but one of those options is better than the other, then the person with the better option has more freedom of choice. Thus someone with *no* choice could have *more* freedom of choice than another!

The second view, associated with Sudgen and Jones (1982), Puppe (1996), and Pattanaik and Xu (1998), is that an additional option has to be at least o.k. to count as freedom-expanding; beyond that options don't expand your freedom just because they would make you better off if you chose them. Of these proposals, that proposed by Jones and Sugden was most influential: they argue that if one of your alternatives is so bad that no reasonable agent could strictly prefer it to your remaining alternatives, then this option cannot contribute to your freedom. Any option which could be reasonably preferred to its alternatives, however, does count as freedom-expanding. This idea avoids the problem with Sen's view, since options are not increasing rewarded for being increasingly desirable, while also capturing to some degree the intuition that expansions in freedom must be associated with expansions in one's ability to live lives one might want to live. Unfortunately, this view has the absurd implication that an individual who receives an additional option that's so good that all reasonable persons must prefer it to its alternatives thereby loses all their freedom.

The problems with both views are ultimately traceable to the fact they analyse option quality in terms of preferences. The responsibility-based approach I defend is a novel interpretation of option quality; it will allow us to construct new measures of freedom which are positively responsive to option quality, are strictly increasing in the number of options, and imply that any two individuals with only one option are equally (un)free. Moreover, the measures I develop are also relatively non-prejudicial in their value assessments.

However, I will ultimately argue that what is actually being measured when we account for option quality in the way I propose is not the extent of people's overall degree of freedom, but rather the intrinsic *value* of freedom, i.e., the value freedom has in itself, over and above its value in enabling us to pursue other ends which we happen to value for independent reasons. Indeed I will argue that my account of option quality connects naturally with arguments for thinking that freedom is intrinsically valuable, and consequently offers an attractive way of measuring how intrinsically valuable each individual's freedom is. So while I take the liberal view that option quality considerations have no role in the measurement of freedom, I make camp with Sen and Jones and Sugden in thinking these considerations have a critical role in explaining why we should care about freedom as an end in itself, and not merely as a means for helping us achieve the aims we care about.

My argument is structured as follows. After briefly motivating Jones and Sugden's preference-based criterion of minimal desirability in section 1, and showing that measures which incorporate this criterion have the troubling implication that more options may mean less freedom, I defend in section 2 a responsibility-based conception of option quality. I discuss two ways of incorporating this new conception of option quality in the measure of freedom: first in the form of a minimal desirability constraint (*à la* Jones and Sugden), and second in a more continuous fashion that recognizes gradations in option quality (*à la* Sen).

In the process of building these measures, I also present a new way of measuring the diversity of a set of options, and fold this into my measures of freedom. Raising the issue of diversity might seem like an unnecessary detour when quality is the topic of the day, but it is in fact necessary to stave off a powerful objection one could otherwise level against my proposed measure. And besides, philosophers and economists often draw a tight connection between the diversity of one's options and their quality, to the point that the two concepts are often cashed out in terms of each other, so it makes sense to comment on how diversity and quality relate to each other on my conception of option quality. The resulting measures are quality-sensitive, diversity-sensitive, strictly increasing in the number of available options, non-prejudicial in their value assessments, and imply that minimal choice implies minimal freedom.

In section 3, however, I argue that while the diversity of one's options is relevant to the degree of freedom they afford us, the quality of one's options is not – this is an unusual but consistent combination of views, given that quality considerations play no role in the measurement of diversity as I interpret it. Consequently, my quality-sensitive measures of freedom are best interpreted as measures of the intrinsic value of freedom, and indeed I argue that they provide very attractive ways of measuring the intrinsic value of freedom.

Throughout this chapter, I bracket accessibility and robustness concerns. I therefore assume the standard setup in which we have a set X of possible options, and the "offers at least as much freedom as" relation, noted $\succsim$, is defined over the set Z of nonempty subsets of X. Opportunity sets are classic sets. In my concluding remarks, however, I will attempt to bring all the considerations we've treated so far together and offer some suggestions as to how they may all be combined.

4.1 Preferences and minimal desirability

Sudgen and Jones (1982) offer the following veil of ignorance type argument for their criterion of minimal desirability: in assessing whether some additional option is desirable or not in some situation, we should abstract from any knowledge we might have about this particular individual's preferences, and ask ourselves instead what, given our general knowledge about human desires, a person similarly situated to this individual might prefer. Through successive application of this procedure we draw up a list of preference orders that reasonable persons in the individual's circumstance might have, and apply a unanimity criterion: an option is not minimally desirable if and only if no reasonable person in this situation could prefer it to its alternatives.

Why take this line? For a start, we may be ignorant about the individual's preferences, and want to avoid prejudicing the criterion of minimal desirability in ways inimical to the individual's actual preferences; veil-based reasoning has the benefit of non-prejudiciality. But more importantly, as Rawls (1971, p. 223) and Scanlon (1998) argue, there is a need in political and moral philosophy to assume a general lack of specific knowledge about individuals' ends, and to admit in our deliberations as permissible any preference profile consistent with a general desire for the sorts of things every person has a reason to value. Imposing any more specific restrictions might license real restrictions on freedom incompatible with the interests of persons.

There are two points to note about this criterion of minimal desirability. First, it is opportunity set relative. In other words, there may be some $x \in X$ and some $A, B \in Z$ such that x is minimally desirable in A but not in B. Second, the notion of reasonableness invoked here is stronger than the standard Humean notion of rationality, which is typically seen to impose only consistency constraints on choice, but not substantive restrictions on what may be preferred to what (Bradley, 2018). Here, though, an option is reasonably preferable to its alternatives just in case a preference for that option would be intelligible to us, and not bizarrely unmotivated, like a general preference for avoiding pain on Tuesdays more than pain on Wednesdays (this example is drawn from Parfit (1981); see also Anscombe, 1957; Brooome, 1999a).

Although Jones and Sugden were the first to introduce this notion of minimal desirability, they did not take the next logical step of devising a ranking rule over Z which builds on this criterion. Pattanaik and Xu (1998) do, however. Define max$A$ as the set of all elements in A that are ranked first by some preference order that a reasonable person may have, then Pattanaik and Xu propose: $A \succsim B \iff |\max A| \geq |\max B|$.

At first blush, this rule seems attractive. If no reasonable person could prefer some options to their alternatives, you might think they're so valueless it doesn't make sense to include them in the measure of freedom – it would be like counting monopoly dollars as part of your income – and therefore what this measure says is that you are freer than I am iff there's more you are free to do, ignoring the options that don't count. Unfortunately, this rule has the odd implication that if you start with n options, each of which could be preferred to its alternative by some reasonable person, and you then receive some additional option which strictly dominates it under every reasonable preference ordering, then you must have strictly less freedom now than before. This seems wrong. Giving people valuable new opportunities while leaving the old ones intact doesn't make them *less* free to choose what to do or how to live. It may make the choice less *interesting*, insofar as if it is a no-brainer that you ought to choose x, then you don't have to engage any of the critical faculties of discrimination, sensitivity, and judgement that ordinarily contribute to making the act of choice worthwhile. But it does not follow from the fact that an expansion of the choices available to one makes choice less interesting that it diminishes one's freedom.[1]

4.2 A responsibility-based account of quality

What drives this unfortunate result is that Jones and Sugden's criterion of minimal desirability relativizes desirability to the opportunity set. This is not surprising, since the standard of minimal desirability is defined in terms of preference, which is a comparative evaluative attitude. Preferences don't admit of good and bad options, only better and worse ones, and so do not allow for a standard of minimal desirability that applies across opportunity sets. But this is exactly the kind of standard we require to avoid the pitfall of Pattanaik and Xu's view while still discounting "valueless" options. I propose that instead of preference, we reach for a noncomparative evaluative attitude, such as approval, to define the standard of minimal desirability.

[1] Xu and Puppe (2000) propose a different rule that ranks sets by their share of maximal elements in their union, which solves the problem highlighted here. However, this ranking is not in general transitive, so I will not discuss it further.

Approval is a concept that appears in the literature on voting systems (e.g., Brams and Fishburn, 1978; Merrill, 1979). It is a subjective, non-comparative evaluative attitude to the set of candidates (in our case, options) which divides it into two subsets – that of candidates the subject approves of, and that of candidates they disapprove of (note that one or the other subset could be empty). What makes approval noncomparative is that whether some option meets the standard is invariant to what alternatives it is being compared to. The standard itself, however, may be either comparative (e.g., approve of x iff x is close enough in value to some top possible alternative) or noncomparative (e.g., approve of x iff x commits no injustice). It is related to preference in that what is approved of must be preferred to what is not approved of.

I propose that an option is approval-worthy if and only if it is the sort that always provides standing for responsibility when it is an available alternative in a non-singleton opportunity set. Pettit and Hoekstra (2018, p. 78) argue that our practices of holding each other responsible – in blaming wrongdoing and praising good action – are sensitive to the number and quality of people's options. We hold responsible only those who *had* a choice about what to do, unless they were also antecedently responsible for their lack of choice; also we don't tend to blame people for wronging others when their only alternative was to do something which would have been overly costly to themselves. Whether some option is too costly or not can depend on context: you may be excused for telling a white lie when you stand to have your arm broken for telling the truth, but you are not excused from murder when you stand to have your arm broken for refraining. At the same time, some options are attractive enough in the eyes of society, or at least so trivially costly, that they could never excuse even minor wrongs: if your only alternative to doing x is to enjoy a pleasant all-expenses paid vacation to New Zealand, then you can be held responsible for x-ing, no matter what x-ing is. Our practices of holding each other responsible implicitly establish a division in the space of options between those options which are "good enough" to always provide standing for responsibility, and those which are not.

Observe that the existence of "good enough" options follows from the assumptions that (1) X is finite and that (2) an option provides standing for responsibility if and only if, according to common standards of assessment, x is not horribly worse than its alternatives. This is because, if we represent common standards of assessment by a (not necessarily complete) weak preference relation R defined over X, it is a consequence of the fact that X is finite that a maximal set of alternatives $\mathcal{X} = \{x \in X | \forall y \in X, \neg yPx\}$ exists. No $x \in \mathcal{X}$ will ever be worse, let alone horribly worse, than any of its alternatives, so it must always be responsibility conferring. Also, any option that is not horribly worse than any of the elements in $\mathcal{X}$ will necessarily also count as good enough.[2]

The standard of approval being suggested here is of normative interest: it is important for individuals to have the standing to be held responsible. Fitness for responsibility is what allows us to enter into other kinds of morally valuable relations, like friendship and contracts, and to express certain values like love and care. Love, characteristically, can only be expressed through acts which one has freely chosen over alternatives which are good enough that one could have been held responsible for choosing them instead of performing the act of love. I don't show you love by whisking you off to a romantic getaway if I was threatened at gunpoint to do it, but I do show you love by whisking you off to a romantic getaway if I could have gone fishing with the guys instead. This criterion doesn't assume a lack of knowledge about individuals, but nor should it: it is the standing to be held responsible in society as it actually is, not as it might be behind the veil of ignorance, which matters to us. Moreover, note that because shared practices of holding each other responsible presuppose common standards of assessment, the value assessments which underpin our practices may safely be assumed to be widely shared. Thus, if one were compelled by the thought that the extent of freedom depends partly on the quality of the options, but wanted to remain non-prejudicial in one's value-assessments, a plausible view to adopt would be that approval-worthy (interpreted here as meaning "responsibility-conferring") options expand individual freedom more than disapproval-worthy ones.

Having established approval as the appropriate standard, let us define $\mathcal{M} \subset X$ as the set of all minimally desirable alternatives in X. Consider now the following conditions on $\succsim$:

(INS) Indifference between no-choice situations: $\forall x, y \in X,\ \{x\} \sim \{y\}$.

(SM) Strict Monotonicity: $\forall x, y \in X,\ \{x, y\} \succ \{x\}$.

(MM) M-Monotonicity: $\forall A, B \in Z$, if $|A \cap \mathcal{M}| = 2$ and $|B \cap \mathcal{M}| = 1$ then $A \succ B$.

(S) Stability: $\forall A, B \in Z$, and $\forall x, y \in X$ such that either $x, y \in \mathcal{M}$ or $x, y \notin \mathcal{M}$, $A \cup \{x\} \succsim B \cup \{y\} \iff A \succsim B$.

(SM) is a very weak version of the principle that more choice means more freedom – it just requires that *some* choice implies more freedom than *no* choice.

[2] Even if X were infinite, the existence of good enough options would be guaranteed provided there is a least upper bound on how good an option can be, which is plausible enough.

This applies even when one or both of the options fails to pass the standard of approval, because even if the additional option doesn't always give standing for responsibility it does give you *some* freedom of choice. Any plausible measure of freedom should imply (SM). (MM), meanwhile, requires that a person with two minimally desirable alternatives has more freedom of choice than a person with a single minimally desirable alternative, regardless of how many other options they both have. This just encodes the key intuition behind minimal-desirability-based measures that options which don't pass the minimal threshold are of negligible value.

(INS) seems like a *sine qua non* condition on a measure of freedom. If two people both have no choice about what to do, then neither can be said to enjoy more freedom than the other. Perhaps surprisingly, this condition is highly contested. Notably, Steiner (1983), Carter (1999), Carter (2004), and Kramer (2003) argue that your freedom is not merely increasing in the number of actions you can pursue, it is also strictly *decreasing* in the number of actions you are physically prevented from doing. (INS) rules this out, because if you and I are both forced to do x, but if forcing you to do x *ipso facto* prevents you from performing more acts than forcing me to do x does (e.g., because I'm disabled), then (INS) ranks us as equally (un)free, but Steiner & co. would rate you as less free. I'm inclined to regard this as a *reductio* against their view. An able-bodied slave and a slightly less able-bodied slave, both forced to slave without any choice about anything, strike me as equally unfree. The able-bodied slave may be *deprived* of more freedom, but at the present time both slaves enjoy equally little freedom.

A different sort of objection to (INS) comes to us from Amartya Sen (1990c, pp. 470-471), who argues, à la Hobbes (1994, L: 79), that a person who is forced to do what they anyways prefer to do and would have done without compulsion (e.g., walk home normally) is freer than someone who is forced to do what they would rather not (e.g., walk home by hopping on one leg). However, it seems to me that what we have here is a conception of voluntariness, not freedom. It is a fairly plausible view of voluntariness on which an action is done voluntarily just in case it is the act one would have done had one been free to choose to do otherwise. But note that this view presupposes a notion of freedom on which freedom requires actually having choice (for a defense of this view, see Oppenheim, 1981). Indeed Sen's own explanation crucially appeals to the fact that "I would choose to walk given the choice." But then I am inclined to say that in Sen's example the subject is not free to choose how to go home in either situation, although in the second case she voluntarily returns home in the way she was forced to. These considerations justify retaining (INS).

That leaves (S), which requires that if you and I both receive an additional option that is minimally valuable, or both receive an option that is not minimally valuable, then I now enjoy at least as much freedom as you do just in case I already enjoyed at least as much freedom as you did. This condition, unfortunately, is not very attractive: if x is very dissimilar to my existing alternatives, but quite similar to your existing alternatives, it is plausible that x makes a greater contribution to my freedom than to yours (G A Cohen, 1978; Arneson, 1985; Sen, 1990c; Pattanaik and Xu, 2000; Dowding and Van Hees, 2009). To see this clearly, note that I seem to have greater freedom of choice if I have the freedom to buy any one of ten differently-coloured but otherwise identical sedans than if I merely have the freedom to purchase any one of ten completely identical sedans. In both cases, I have equally many options to choose from, but intuitively I have more to choose between in the first case than in the second, since I can choose the colour of my car. One could, I grant, insist that while the quality of my options matters, their diversity does not, but this would be odd because the whole motivation for getting away from just counting how many options people have was to arrive at a measure of freedom that recovers more intuitive data. Accordingly, the next section develops ranking rules which are sensitive *both* to quality *and* dissimilarity concerns, and therefore violate (S), but for now, before we add any more bells and whistles, I think it is worth getting clear on the logic of how a ranking rule can register concerns for minimal desirability, in isolation of other considerations. So consider:

Theorem 4.2.1. $\succsim$ *satisfies (INS), (SM), (MM), and (S) iff, for any two* $A, B \in Z$*:* $A \succ B \iff max\{1, |A \cap \mathcal{M}|\} > max\{1, |B \cap \mathcal{M}|\}$ *or* $(max\{1, |A \cap \mathcal{M}|\} = |max\{1, |B \cap \mathcal{M}|\}$ *and* $|A| > |B|)$*. And moreover* $A \sim B \iff max\{1, |A \cap \mathcal{M}|\} = max\{1, |B \cap \mathcal{M}|\}$ *and* $|A| = |B|$

The proof is in the appendix. This ranking rule orders opportunity sets via a two-step procedure: sets are first ordered by the maximum over 1 and the number of minimally valuable alternatives they contain. In the case of ties, they are then ranked by the total number of options they contain. Note that if we simply ordered opportunity sets by the number of minimally desirable alternatives they contained, and then resolved ties by counting the number of remaining options, that (INS) would be violated. This rule has the attractive implication that receiving an additional option always expands your freedom. It is sensitive to the thought that options which pass minimal standards of review expand your freedom more than options which do not – to the point that no additional number of not-minimally valuable options can make up for the loss of a single minimally valuable option. It's also a very tractable measure: all we need in order to apply it is to be able to distinguish between options that are always responsibility conferring and options that aren't. These are fairly coarse-grained judgements, so less information is needed to reach them, and we can hope they will be fairly widely shared, since, being coarse-grained, they will be less sensitive to differences in individual dispositions to make judgements. If we can further sensitize this measure to diversity concerns, we'll have a decent candidate for a measure of freedom. This is the task to which I now turn.

4.2.1 Incorporating diversity

We start with a dissimilarity function $\delta : X \times X \longrightarrow \mathbb{R}^+$ which orders pairs of options by how dissimilar they are to each other. Possible interpretations of δ are that it orders pairs of options by (1) how likely any given individual is to fail to be indifferent between them, (2) how intensely any given individual is expected to prefer one to the other, or (3) how similar the possible world in which you exercise one option is from the possible world in which you exercise the other. Next, we define a function $d : X \longrightarrow \mathbb{R}^n$ ordering all the elements in X according to how dissimilar they are to some arbitrary $x_0 \in X$, chosen as reference, along each of the n possible dimensions of dissimilarity. Thus d associates to every possible option an n-tuple (i.e., a row of n numbers) which define its coordinates in a Cartesian grid of origin x_0, and the only constraint we impose on d is that the euclidean distance between any two options x, y equal $\delta(x, y)$. Thus what d gives us is really a geometric representation of option dissimilarity, where options are represented as points in an n-dimensional space, and spatial relations between these points represent dissimilarity relations.

This construction makes sense if (but only if) we assume that dissimilarity is cardinal, i.e., we can meaningfully compare magnitudes of differences in dissimilarity between pairs of options. This is reasonable on each of the interpretations of d I've suggested. Probabilities of failing to be indifferent between any two options are cardinally measurable. More controversially, many have argued that intensities of preference satisfaction are as well (Adler, 2011; Greaves, 2017), and I argued in the previous chapter that world-similarity is too (see also Krodel and Huber, 2013). Note that if we interpret dissimilarity in terms of preference, then the n dimensions of dissimilarity can be interpreted as the n arguments of people's utility functions, while if we interpret dissimilarity in terms of similarity between possible worlds, then they are naturally interpreted as the n dimensions along which world-similarity varies.

It is worth highlighting that dissimilarity, as I am thinking of it here, is not a qualitative concept. This is important because many philosophers, and most notably Garnett (2016) and Carter (1999, p. 121), argue that any assessment of dissimilarity must be value-laden, because it must presuppose a point of view on which differences among options *matter*. Not so. None of the three interpretations of option-dissimilarity that I suggested in the previous section require evaluative judgements. The two preferencist interpretations are grounded entirely in descriptive facts about how likely individuals are to prefer one option to another, and how intensely. These measures certainly assume a particular picture of what dissimilarity between options *is*, namely, the propensity to provoke divergent evaluative responses. But this is a metaphysical judgement, not an evaluative one. Likewise world-similarity, which denotes the extent to which two possible worlds "agree" about the way things are, is determined by descriptive facts about which counterfactuals are true, at least on the standard Lewisian model of counterfactual semantics (Lewis, 1979).

We're now in a position to propose a philosophically well-grounded diversity function. Very roughly, we would like a measure of diversity to reward opportunity sets for having more options, but also for "dispersing" their options more widely over $\mathbb{R}^n$. There's an elegant way to do both. We associate to every possible option an n-dimensional Gaussian function – i.e., a *bell curve* that reaches its peak at the exact point where the option is located in $\mathbb{R}^n$ – and we then take the integral over $\mathbb{R}^n$ of the maximum value assigned to any point by the family of Gaussian functions.

To visualize this clearly, imagine nailing an elastic sheet over a flat surface, and then deforming the sheet by placing narrow pegs underneath it; locally, the sheet will reach a peak wherever we place a peg, then flatten out as we move away from it (though without ever becoming completely flat). A single peg thus turns our sheet into a two-dimensional bell curve. Now clearly, if you place more pegs underneath the sheet, you deform it more. Moreover, the further you drag these pegs away from each other, the sheet becomes less and less flat over more of its surface, reaching multiple peaks in multiple places, and only gradually flattening out. In this image, pegs represent the coordinates of options in $\mathbb{R}^n$, and the empty space between the sheet and the surface it's nailed to is our integral.

We can state this all precisely. Let $g : \mathbb{R} \to \mathbb{R}$ be the function $g(t) = e^{-t^2}$. Fix an integer $n \geq 1$. Given $z \in X$ and $d(z) \in \mathbb{R}^n$, let $g_z : \mathbb{R}^n \to \mathbb{R}$ be the function $g_z(x) = g(\|x - d(z)\|)$.

We define the diversity function $D : Z \to \mathbb{R}$ by setting, for any $A = \{a_1, ..., a_k\}$,

$$D(A) = \int_{\mathbb{R}^n} \max\left(g_{a_1}, \ldots, g_{a_k}\right).$$

D is a novel way of measuring diversity.[3] And it can be shown (see appendix) that this diversity function satisfies the four following attractive and philosophically significant conditions.

> **(M) Monotonicity**: if $A \subset B$ then $D(A) < D(B)$.
>
> **(D) Dispersion**: for any set A and any option $x \in A$ such that $d(x)$ does not lie within the convex hull $\mathcal{K}$ of A, if $\mathcal{H}$ is a hyperplane that does not intersect the interior of $\mathcal{K}$, then $D(A \cup \{x\})$ strictly increases as $d(x)$ moves away from the convex hull of A along a line that is orthogonal to $\mathcal{H}$.
>
> **(C) Continuity**: if x, y are two elements in A then $lim_{d(x) \longrightarrow d(y)} D(A) = D(A \setminus \{x\})$.
>
> **(TI) Translation Invariance**: if there is a bijection $f : A \to B$ such that for all $x \in A$, $d(f(x)) = d(x) + c$, where c is some constant vector, then $D(A) = D(B)$.

(M) says that each additional option you receive strictly increases the diversity of your choices. This is attractive, since you retain all your previous options (so your opportunities can't well now be *less* diverse), but you've now got some new option that differs in some way from those previously available alternatives.

(D) is a condition that rewards opportunity sets whose elements are more dispersed over $\mathbb{R}^n$. Very roughly, it requires that as you "drag" one point from some set further and further away from the clump where the other points in the set are located, the set of points becomes more dispersed over $\mathcal{R}^n$. We can visualize this very clearly in one dimension: suppose you plot two bell curves on a graph, both with the same variance and integral (i.e., their peak is at the same height and the area under each curve is the same) but slightly different mean values (i.e., they reach their maximum at different albeit nearby points). The two bell curves therefore overlap with each other over most of their area, which shows that their peaks are not very dispersed over the real line. But now imagine that I "pull" the peak of one of the two bell curves away from the other. Then the total area of overlap between the two bell curves will diminish, and they will jointly cover a larger area, mirroring the fact that their peaks are now more dispersed over the real line.

[3] See Weitzman (1992), Pattanaik and Xu (2000), William Bossert, Pattanaik, and Xu (1992), Baumgärtner (2006), Bervoets and Gravel (2007), Pattanaik and Xu (2008), Hees (2004), and Nehring and Puppe (2009), for alternative approaches.

(C) and (TI), finally, are mostly technical properties of F, but they are nonetheless plausible and have philosophically significant interpretations. (C) requires that an option which differs from another alternative only to an infinitesimally small extent can only make an infinitesimally small marginal contribution to the diversity of your opportunity set. This condition is attractive for two reasons: first, continuity is generally an attractive condition of value functions, because this is necessary for the existence of solutions to optimization problems. If, say, we care both about individual diversity of opportunity and individual happiness, but those two goals compete to some extent and we have only finite resources with which to pursue them, then, although we may disagree on the optimal allocation of resources, we can at least guarantee the existence of *efficient* allocations if diversity and happiness are both measurable by continuous functions. Second, it's just intuitive that insignificant differences between options provide only insignificant diversity of choice. To take an example from Sugden (1992, p. 318), suppose you're offered a choice between one of three cans of the exact same beer on a flight, each different from the other only by their serial number and position on the tray; you have *some* diversity of choice in this situation (choosing the can on the left is not *quite* the same as choosing the can on the right), but so vanishingly little one is almost tempted to say you only really have one option before you. Contrast this with a situation where you can choose between a can of beer, a small bottle of wine, and water.

(TI), finally, essentially just states that if the options in A are as dissimilar, relative to each other, as the options in B, relative to each other, then A and B must be equally diverse. Or in other words, an option's absolute position in $\mathbb{R}^n$ has no bearing on how great a contribution it makes to the diversity of your options: only its relative position to available alternatives matters. This condition is attractive, again, for two reasons. First, our choice of origin for $d(.)$ was arbitrary, but the diversity of a collection supervenes on the elements of that collection. A measure of diversity whose assessments depended on a conventional choice of origin for $d(.)$ would be completely pathological, as it would imply that the diversity of a collection can change without the elements in that collection changing. Second, diversity is a property of the *relations* in dissimilarity between options, not a property of the *intrinsic* features of options. Thus if you change the intrinsic properties of a set of options, but preserve all the relational properties of dissimilarity, you can't have changed the overall degree of diversity in the set. This is most evident when we consider two singleton sets. Any option x is as similar to itself as any other option y is similar to *it*self, so with no other basis for making judgements of diversity, we're left to conclude that $\{x\}$ and $\{y\}$ are necessarily equally (un)diverse.

D does not satisfy these conditions uniquely (one didn't need to use Gaussian functions, for instance), but it is, I believe, the first measure of diversity yet proposed to satisfy them all. Moreover, there is an informal argument for measuring diversity using D. One way of interpreting the Gaussian functions is as a representation of substitution value. If x and y are very similar, then they must be good substitutes for each other: if you can't do y (e.g., eat some particular apple) but you can do x (e.g., eat a different apple), it's almost as if you could do y. Notice that in this case $d(x)$ and $d(y)$ must be very nearby in R^n, so the Gaussian associated with $d(x)$ will reach a high value at $d(y)$. In contrast, if x and y are very dissimilar, then the Gaussian associated with $d(x)$ will reach a very low value at $d(y)$, mirroring the fact that x and y are very complimentary. So the Gaussian associated with $d(x)$ distributes its density in a way that tracks how good a substitute x is for other possible options. This interpretation gives us a very natural way of thinking about diversity: a very diverse opportunity set, as measured by D, is a set that maximizes the space of possible options for which the options that are *actually* available are good substitutes. In fact, not only is this a natural way of thinking about diversity, it helps explain why diversity matters. If x and y are near-perfect substitutes, then having x gives one a sort of proxy access to y, even if you don't have y. In effect, having more diverse options is a way of having access to more options – without *actually* having more options.

In sum, D is an attractive way of measuring diversity in $\mathbb{R}^n$. Now consider the following ranking rules.

(a) $\forall A, B \in Z, A \succsim B \iff D(A) \geq D(B)$

(b) $\forall A, B \in Z, A \succ B \iff$ either $max\{1, D(A \cap \mathcal{M})\} > max\{1, D(B \cap \mathcal{M})\}$ or $D(A \cap \mathcal{M}) = D(B \cap \mathcal{M})$ and $D(A) > D(B)$. Moreover $A \sim B \iff D(A \cap \mathcal{M}) = D(B \cap \mathcal{M})$ and $D(A) = D(B)$.

Rule (a) simply states that the degree of freedom offered by opportunity sets is increasing in the diversity between the options in that set. This rule is of interest to anyone who wishes to sensitize their measure of freedom to dissimilarity considerations, but not to option quality considerations. We'll come back to it later. Rule (b) is a version of our earlier rule, now sensitized to dissimilarity conditions. Rule (b) retains all the attractions of our earlier lexical rule, while avoiding the objection that it ignores diversity considerations in measuring individual freedom. Therefore, if one is to incorporate a concern for option quality in the measure of freedom, and if this concern is to be incorporated by a minimal desirability constraint, then rule (b) is an attractive solution concept.

4.2.2 From a lexical rule to a continuous rule

That said, one might want to incorporate qualitative constraints on the measure of freedom yet resist the use of a lexical rule. After all, the capacity of an option to excuse wrongdoing due to its costly character is graded, so if an option's quality is determined by its tendency to excuse wrongdoing, a bivalent standard like approval is too crude a measure of quality.[4] And anyway, whatever gains in tractability applying a lexical measure may afford us, it is not plausible that options which are guaranteed to provide standing for responsibility are lexically superior in value to options which are not. It would imply that we should either not be willing to pay anything for options that don't pass minimal standards of review, or that we should be willing to pay any price for options that do. The latter disjunct is obviously false, and the former is also implausible: even terrible options will provide standing for responsibility in some cases, and so would be worth paying something for.

To address these concerns, we can try extending rule (b) to a continuous order. Consider that if an option would confer responsibility to a given individual across more possible worlds (within a class Ω of possible worlds that taken as relevant to this assessment) than another alternative, then the first option is in an obvious sense more capable of grounding standing for responsibility, and for that reason more valuable to the individual. Now, since the set of possible worlds is representable as a Euclidean space, we can in principle assign to sets of possible worlds a number representing how big they are. In particular, if we impose an upper bound on Ω, limiting our attention only to those worlds within some sufficient distance from the actual world, then we can assign numbers between 0 and 1 to any subset of Ω denoting how large a share of Ω the subset occupies. Now assume that an option's quality is linearly increasing in the share of situations across which it is responsibility-conferring ("good enough" options get a value of 1, because they're always responsibility-conferring). Then there exists a value function $v : X \longrightarrow [0, 1]$ representing option quality. Now consider the following ranking rule:

$$A \succsim B \iff (D(A) - 1) \cdot \sum_{x \in A} v(x) \geq (D(B) - 1) \cdot \sum_{x \in B} v(x).$$

Call this the continuous quality-based rule. This rule is a simple extension of rule (b) which satisfies (INS) because $D(A) - 1 = 0$ whenever A is a singleton, implies that expansions in one's opportunities always increase our freedom, and

[4] Incidentally, the reasonableness of a preference order is also graded: behind some veil of ignorance, some preference orders will seem more plausible from behind the veil as candidates for people's actual preferences, so the present criticism applies to Jones and Sugden's criterion as well.

implies that the degree of freedom offered by an opportunity set is strictly increasing in both the value of one's choices and in their degree of diversity. Thus the continuous quality-based rule avoids the pitfalls associated with Sen's view, which violates (INS), while capturing his general intuition that freedom expands with our capacity to live lives that are of better quality (according to common standards of assessment); it is sensitive to the fact that option quality is graded, and it implies no implausible lexical superiorities in value.

4.3 Responsibility and the intrinsic value of freedom

With all that said, I will nonetheless argue in this section that option quality considerations shouldn't be taken into account in the measurement of freedom, having a more useful role in measuring the *value* of freedom.

There is an interesting contrast between the sorts of reasons appealed to by philosophers who argue that freedom is increasing in the dissimilarity between one's options, and those appealed to by philosophers who argue that freedom is increasing in the value of one's options.[5] Both groups agree that merely counting options is silly but, as stated earlier, the intuition behind why dissimilarity matters is that a set of highly dissimilar options gives you *more to choose between* than a more homogeneous set. Indeed, there is a quasi-literal sense in which having more diverse options gives you more options. In contrast, the intuition behind why quality matters is that better options enable people to live better lives – and insofar as we care about people's freedom, don't we care about their freedom to live well?[6] To which the answer is obviously yes. We're only in the game of measuring freedom because we think it's valuable, in both intrinsic and non-intrinsic ways. But to show that we only care about measuring freedom because it is a valuable thing is not to show that the measure of freedom itself should be value-laden.

And indeed, to the contrary, a good measure of freedom should only register how extensive one's opportunities for choice are, not how good these are. This is because a measure of freedom ought to be neutral on the question of whether more freedom is (in any respect) better than less. A measure of freedom that rewards opportunity sets for enabling individuals to live better lives is not neutral on this question. It is not a conceptual necessity that more freedom is (*ceteris paribus*) better than less – I happen to think that it is, but this is a substantive,

[5] Of course, many defend both views simultaneously, e.g., G A Cohen, 1978, §5.4, Raz, 1988; Sen, 1990c.

[6] See Raz (1988, p. 374) and Sen (1990c, p. 471).

disputed claim; if we do not wish to speak past each other we must first agree on a way of measuring what counts as "more" or "less" freedom that does not prejudice the outcome of our dispute. Moreover, I share the standard liberal attitude that our conception of freedom ought to be nonprejudicial about what ends are worthwhile, for the familiar reasons that one's own value judgements may be wrong, that some disagreement about ends may be faultless, and that restrictions on people's freedom must be justifiable in terms acceptable to all, etc. (see Garnett, 2016). In sum, while I think it *is* appropriate to sensitize one's measure of freedom to diversity considerations, it is inappropriate to sensitize it to quality considerations. Rule (a), therefore, is a better measure of freedom than the continuous quality-based rule.

Moreover, I submit that option quality considerations are simply more illuminatingly invoked to explain what makes freedom valuable than to explain what makes for more or less freedom. My responsibility-based account of option quality is well-suited to explain the *intrinsic* character of freedom's value, that is, the fact that freedom may be valuable in itself, in a way that entails that more freedom is (*ceteris paribus*) better than less, and in a way that does not reduce to its instrumental value in enabling us to pursue further ends that we value independently. This is in contrast to other quality-sensitive measures of freedom, such as have been proposed by Kreps (1979), Arrow (1995), and Vallentyne (2002), which are ill-suited to explain the intrinsic value of freedom, because they essentially reduce the value of freedom to its expected utility, which is not necessarily increasing in one's freedom.

If freedom is intrinsically valuable, then more freedom is, *ceteris paribus*, better than less. This implies that it is sometimes better to have a choice between some set of options A than to simply be forced to do x, where $x \notin A$ and you prefer x to any $y \in A$, because the intrinsically valuable gain in freedom will sometimes offset the intrinsically disvaluable loss of well-being (G. Dworkin, 1982, p. 60). Hurka (2011, pp. 142-143) persuasively defends this implication on the grounds that in the first case, but not in the second, you are responsible for what you do and bring about, and have a more active role in shaping your own life.[7] Living a life of one's own making is simply a valuable component of living a good life, valuable enough that it is worth not always getting what one wants in order to be a person of one's own making. Or to put things differently, freedom is intrinsically valuable because being free is a valuable way to be. It comes with a unique moral status: the standing to say "this is me, this is what I choose, this is how I live, this is what I value." Moreover, Hurka goes on, given that one is more responsible

[7] This is a point also emphasized by Sugden (1992) and Arrow (1995).

for the shape of one's life the more ways there are one could have chosen to live it instead, it follows that that more freedom is (*ceteris paribus*) better – up to, I would stress, how good additional options are at conferring on us standing for responsibility.

This last point by Hurka, qualified in the way I propose, speaks strongly in favour of the continuous quality-based rule as a measure of the intrinsic value of freedom, under a responsibility-based interpretation of option quality. Interpreted in this way, the rule states that the intrinsic value of freedom of choice is increasing in the extent of one's freedom, and in the capacity of one's available options to ground standing for responsibility, which is just a formal explication of Hurka's intuition. The rule also has the desirable implication that when one has minimal freedom of choice (i.e., one's opportunity set is a singleton) then the intrinsic value of the freedom one enjoys is minimal, which is what you would expect on a view like Hurka's. If my opportunity set is a singleton, i.e., I can only do one thing, then I'm not responsible for how my life then turns out, and I don't enjoy the status of a free person, i.e., one who can choose how to live according to their will and be responsible for it. If it's this status that is valuable, then the degree of freedom I currently enjoy couldn't be less intrinsically valuable. Further, this rule in unresponsive to individual preferences over their options, which implies that the intrinsic value of your freedom is invariant to whether you happen to value the ends your freedom allows you to pursue. This is it as it should be: the intrinsic value of freedom cannot depend on its instrumental value in satisfying your desires (including your desire for more freedom or responsibility over your life). And finally, this rule has important formal properties, such as continuity, and the fact that it trivially induces a transitive value ordering over Z. These are all attractive properties for a measure of the intrinsic value of freedom.

Naturally, in offering the continuous quality-based rule as a measure of the intrinsic value of freedom, I am committing myself to the claim that freedom has intrinsic value. Hurka's argument notwithstanding, many philosophers balk at the claim that it can be better to be free and not get what you want than simply to get what you really want (G. Dworkin, 1982; Gustafsson, 2019), and so have offered skeptical arguments in response. The most recent attempt to refute the claim that freedom is intrinsically valuable comes from Gustafsson (2019), and so, in closing, I think it is worth addressing his worries.

Gustafsson's argument can be simplified as follows: if we receive additional options which any rational person is required to strictly disprefer to their existing alternatives (what he calls a valueless expansion), then there is no intrinsic value in the expansion in our freedom – it is *not*, all else being equal, better to have those additional options than not to have them. Freedom is therefore not intrinsically valuable, because if freedom is intrinsically valuable then there is intrinsic value in every expansion in our freedom, including valueless expansions.

Gustafsson (2019, p. 10) defends this claim by imagining that the valueless expansion happens in two steps. Suppose that you start with a single option. In the first step, you're to imagine you receive two additional options which are exactly alike to the first in every respect that a rational person is permitted to care about. So, for example, you start with an orange, and are then given a choice between that orange or two additional oranges that are physical duplicates of the first. After this first step, claims Gustafsson, you have not gained anything of intrinsic value, because the new opportunities you've acquired don't differ from the starting option in a way that it is rational to care about – and if you *had* gained something of intrinsic value it should at least be rationally permissible to prefer the new situation to the old one. In the second step, you imagine each of the two additional options being worsened in some distinct way. You now have three options, two of which are strictly worse than the first, e.g., one of your oranges rots and the other one dries out. The deterioration can't have improved the intrinsic value of your freedom of choice, says Gustafsson, because it leaves you with as many options as before, and just makes some of your options worse without improving any in return.

There is much to unpack in this argument. For a start, it is important to underscore that Gustafsson is presupposing a non-Humean conception of rationality, according to which there are rational requirements on the content of preferences. There are thus some things which it is irrational to prefer to others, even if this implies no inconsistency in one's choice pattern. Gustafsson's argument is therefore easily resisted by insisting on Humeanism, which is the prevailing view of rationality within economics and philosophy. Humeanism also fits naturally within a liberal worldview, since liberals are hostile to people telling others what they can or cannot like. Carter (1999, pp. 54-99), for example, implicitly invokes Humeanism to reject quality constraints on the measure of freedom, arguing that no one is in a position to declare that some option just couldn't be valued by someone, or that some aim could not be rationally pursued. I agree. Still, I will grant Gustafsson the conception of rationality he presupposes, and respond directly to his two-step argument.

Both steps in this argument are dubious. After the first step, you may have very little freedom, but you do have *some*: $\forall x, y, z \in X, D\{x, y, z\} > 0$. In some tiny respect your life will turn out differently depending on what you choose, and you will be responsible for this difference. Being responsible for one's own life is valuable, therefore there is value (however little) in being responsible for this tiny difference. And after the second step, your life could turn out *very* differently depending on what you choose, so you are much more responsible for how your life is lived. Gustafsson might object that if it is irrational to choose anything other than the dominant option, you can't claim too much responsibility for choosing well. But this is false: people *frequently* make irrational decisions, e.g., through displays of loss aversion (Kahneman and Tversky, 1979), or hyperbolic discounting (see Samuelson, 1937; Ainslie, 2001), and they are responsible for those choices as well as for the rational ones. Indeed, you might take great pride in the fact that you're able to avoid making these sorts of errors in judgement. Or, contrariwise, you might revel in making choices derided as irrational just to stick it to the rationalists. After all, as Dostoyevsky (1864, p. 33) wrote in *Notes from the Underground*:

> I, for instance, would not be in the least surprised if all of a sudden, A PROPOS of nothing, in the midst of general prosperity a gentleman with an ignoble, or rather with a reactionary and ironical, countenance were to arise and, putting his arms akimbo, say to us all: "I say, gentleman, hadn't we better kick over the whole show and scatter rationalism to the winds, simply to send these logarithms to the devil, and to enable us to live once more at our own sweet foolish will!" (...) And how do these wiseacres know that man wants a normal, a virtuous choice? What has made them conceive that man must want a rationally advantageous choice? What man wants is simply INDEPENDENT choice, whatever that independence may cost and wherever it may lead. And choice, of course, the devil only knows what choice.

Either way, and whether or not you would actually exercise it, you would value the freedom to make irrational choices. Thus we have ample grounds for rejecting Gustafsson's skeptical argument against the intrinsic value of freedom.

Summary

Where have we come to? I've argued that the most promising kind of quality-sensitive measure of freedom is one based on a responsibility-based account of option quality. Of the two quality-based measures I've proposed, the continuous quality-based rule is particularly attractive, though the lexical rule may have some advantages in tractability. In the process of defending this thesis I have also presented a new and attractive way of measuring the diversity of opportunity sets.

Nonetheless, I've argued that, ultimately, we shouldn't impose quality constraints on the measure of freedom. Rather, these should instead be applied to the measure of the intrinsic value of freedom. And I contend that the continuous quality-based rule is a plausible measure of the intrinsic value of freedom of choice. This measure connects naturally with the best argument we have for thinking that freedom is intrinsically valuable. Rule (a), meanwhile, makes for a more plausible candidate measure of freedom, since it is wholly non-value-laden and strictly increasing in the number and diversity of one's options.

The package of views that emerges from this discussion has an attractive unifying power, because it resolves the question of whether quality considerations are to bear on the measure of freedom in a way that lets everyone in the debate have their cake and eat it too. It gives traditional liberals the victory they wanted: the measure of freedom is insensitive to option quality. But Sen & co. suffer no harm from this defeat, because option quality turns out to be essential to the measure of the intrinsic value of freedom. Thus insofar as we are interested in expanding the scope of individual freedom for its own sake – a central aim of both liberalism and the Sen-inspired capability approach to human development – we will want to refer to quality-sensitive measures.

****

The next chapter will pursue the investigations we've started here into what makes freedom valuable, but before we turn to this, it is time, at last, to pull together everything we've learned so far about the measurement of freedom. How, ultimately, do I think freedom should be measured? Overall assessments of freedom depend on many things: how many options you have, how accessible these options are, how modally robust they are, and how diverse they are. How should one tie all these concerns together? I said in the previous chapter that we don't need to settle on a single measure of freedom: different measures may be appropriate for different contexts. I still think this is a defensible view. For example, in aid and development contexts, it may be best to focus on how many options people have, how diverse they are, and how accessible they are, but ignore modal considerations. Why? Because the first aim of policy should be to make people more capable – to expand the range of things they can do; in any case, we usually can't do much to make people's freedoms more robust in development contexts, since many development initiatives take place under autocratic regimes whose co-operation we need to carry out our interventions, and whose very autocratic nature makes all freedoms substantially non-robust. In contrast, if you're interested in constitutional reform, or if you want to know how worried you should be about corporate or government concentration of power, the modal considerations may be more salient, and accessibility considerations less so.

Still, it may also be useful to have a single measure of overall freedom. And here we can propose two adequacy conditions on how to combine rule (a) with the two fuzzy cardinality rules considered in earlier chapters: first, (INS) must be respected, and second, freedom must be strictly increasing in the number, accessibility, robustness, and diversity of one's options. So consider the triple (A, μ_A, θ_A), where $A \in Z$ is a set of available options, μ_a is a probabilistically interpreted membership function that defines for A an associated fuzzy set $A^* = (X, \mu_A) \in Z^*$, and θ_A is a modally interpreted membership function that defines for A an associated fuzzy set $A' = (X, \theta_A) \in Z^*$. Both θ_A and μ_A have the property that the elements of $X \cap A$ are the only elements to receive a non-zero membership grade. This triple tells us what your options are, how accessible they are, and how robust they are, which is everything we need to know to assess how much freedom you enjoy. So I propose your "overall" degree of freedom is given by a function f from the set of all such triples to the reals. The simplest such function which conforms to our two desiderata is defined by:

$$f(A, \mu_A, \theta_A) = (D(A) - 1)^a \times [\sum_{x \in X} \mu_A(x)]^b \times [\sum_{x \in X} \theta_A(x)]^c,$$

where a, b, c are strictly positive weights reflecting how much importance you wish to give to accessibility, robustness, and variety, relative to each other.

What this rule is telling us, essentially, is that your overall degree of freedom is given by the degree of diversity of your options, discounted by how inaccessible and non-robust your options are. It's easy to verify that this rule has the desired properties. Notice that this rule is not separable across the four factors that determine the extent of your freedom (number, diversity, accessibility, and robustness), but this actually seems right: the extent to which an improvement in any one of these four factors expands your freedom does not seem independent of how well you're doing on the other factors. An increase in the accessibility or robustness of one option expands your freedom less if this is an option that is very similar to many others; conversely an increase in the number or diversity of your options can't be seen to expand your freedom very considerably if the new options are barely at all accessible. Tentatively, then, I am ready to proffer this rule as the correct measure of overall individual freedom.[8]

[8] And obviously, if you add $V(A)^d$ as a fourth term in this product, we will arrive at a complete overall measure of the intrinsic value of the individual's freedom.

Chapter 5

The Diagnostic Value of Freedom

ELizabeth feels conflicted. She has just been notified by email that there are cheap student tickets available to see the ballet version of *Manon* at the Royal Opera House, and also tickets for a production of *Rigoletto*. Elizabeth knows that she likes opera, and she knows that she likes ballet *music*, but she's unsure if she actually enjoys *ballet*, having never gone before. Her preferences between ballet and opera are incomplete: she simply isn't in a position to judge what would make for the best use of her money. In the end, she goes to see *Manon*, where she discovers that she deeply dislikes ballet. In a way, this experience was very valuable to Elizabeth. True, she enjoyed herself less than she would have otherwise, but she acquired valuable information. In the future, if she is ever has to make a decision about whether to go to the opera or the ballet, she will know what she prefers. As I will put it, Elizabeth's choice was *diagnostically valuable* to her.

Elizabeth's case is not uncommon. We frequently elect particular courses of action not because we prefer them to their alternatives, but because we hope to learn something about our own dispositions, tastes, or generally anything that will help us form new or better-informed preferences. Indeed, this may be one of the reasons we care about having a wide range of options available to us in everyday life: if we are free to try out many different activities and pursue different interests, we may discover something about our tastes and desires, or about the features of our alternatives. This in turn may cause us to form new judgments about the good, or revise our previous judgments. Our whole plan of life may be turned upside-down by what we learn. If we were already maximally opinionated, that is, if our preferences were complete and we always had fixed prior judgments about which of two options is better, as is often assumed by rational choice theorists and economists, then having a wide range of options would matter less to us.

This point is under-appreciated by philosophers and economists. Often, in assessing the (instrumental) value of freedom to individuals, theorists have narrowly focused on the extent to which individuals' opportunities enable them to achieve higher levels of welfare. Little concern is expressed for whether individuals have opportunities to learn about themselves and their preferences. This is the issue which this paper addresses: I will be looking at the many different ways in which an option can be diagnostically valuable. My approach is as follows. In section 2, I introduce some of the more influential accounts of the instrumental value of opportunities. I show that a concern for diagnostic value, that is, for the value of learning about oneself and one's preferences, is missing from these accounts. In section 3, I make the intuitive idea of diagnostic value more precise, and briefly contrast my interpretation of diagnostic value with Bodner and Prelec's concept of diagnostic utility. In sections 4 through 8, I discuss the various dimensions of diagnostic value, and I propose ways of measuring diagnostic value along each of these dimensions. Finally in section 9 I consider how to move from these four separate measures to an overall measure of the diagnostic value of a set of options.

5.1 What makes opportunities valuable?

I argued in the previous chapter that freedom is intrinsically valuable. And insofar as freedom is intrinsically valuable, more freedom is better. Of course, freedom is also valuable instrumentally, notably because it allows us to achieve outcomes we prefer. And insofar as we care about its instrumental value, more freedom is not always better. If you dream of being an astronaut, and NASA offers you a position in their space program, an offer from your uncle to take over his second-hand car dealership is of little instrumental value to you. As G. Dworkin (1982) observes, there are decision-making and information-gathering costs to deciding carefully between a large swath of options; if you already have very good options to choose from, giving you a bunch of extra options may be more trouble than it's worth.

How then do we go about assessing the instrumental value of opportunity sets, if we don't simply count up options? A simple approach would be to look at individual preferences over options. Pattanaik and Xu (2015) suggest that we could order opportunity sets by their top-ranked option. That is, an opportunity set A is instrumentally better for you than another set B just in case you prefer your top-ranked option in A to your top-ranked option in B. G. Dworkin (1982, p. 60) also seems to endorse this view, arguing that if a, b, c are three objects, which I rank in that order, then it is better for me to receive a than to be given a choice between b and c. If, moreover, intensities of preferences are interpersonally comparable, then we can say that opportunity set A is more valuable to Stephanie than set B is to Delilah just in case Stephanie's desire for her top-ranked option is more intense than Delilah's desire for hers. Thus, on this model, an additional option is valuable to you just in case you strictly prefer it to every one of your other alternatives (note that this view is insensitive to the problem of decision-making costs).

Now, looking only at top-ranked options is surely naïve. If one's preferences are incomplete, and we have no top-ranked options, we may prefer having all our "maximal" (i.e., undominated) options to only one of our maximal options. After all, even if today I'm torn about whether I want to go see *Götterdämmerung* or *Hamlet* and just can't decide how I want to spend my money, tomorrow I might make up my mind – indeed, I might rationally be willing to pay £10 for the guarantee that there will still be tickets for both performances until tomorrow evening, just to keep my options open. More generally, as Kreps (1979), Sen (1993b), and Pattanaik and Xu (2015) point out, if we're unsure what our future preferences will be, then we may want more options than just those few which we currently most prefer. If I know myself to be capricious (or, more kindly, spontaneous), I may anticipate that my tastes will change abruptly, and thus although I *currently* rank x over all of its alternatives, I may prefer having every alternative available to me to just having x.

Further, we may place what Scanlon (1998, pp. 252-253) calls *representative* and *symbolic* value on having certain options. Options have representative value if choosing between those options enables us to express our values. Scanlon gives the example of selecting a gift for one's partner. In these cases, we don't simply care about securing the best possible gift for our partner (though we *do* care about this); we also care about choosing it ourselves, after careful thought, from among a set of alternatives, because it is only through this sort of careful choice that we can express our love and commitment to them. If an omniscient third party comes along and threatens to present us with the gift that our partner would most enjoy unless we pay them, it may be perfectly rational to pay to retain the privilege of making our own selection.

Meanwhile, options have symbolic value if *not* having those options to choose from "would be seen as reflecting a judgment (their own or someone else's) that they are not competent or do not have the standing normally accorded an adult member of the society" (Scanlon, 1998, p. 253) . Thus if we care about being seen as an individual capable of individual choice, if we refuse to be coddled by a parternalist entity, we may rationally prefer to have many options to having just the one option we do (or would) most prefer. And indeed, we may be willing to retain the services of a protective agency whose job it is to guard us from such intervention.

The lesson here seems to be that we need to attend to individuals' global preferences for sets of options in assessing the value of a set of alternatives to them, not merely their preferences over individual options. We could thus simply say that A is instrumentally better for you than B just in case you rank A over B, where this ranking reflect your preferences over options, your interest in keeping your options open, and your interest in having representatively and symbolically valuable options. Or else, as Pattanaik and Xu (2015) suggest, that A is better for you than B just in case you rank the pair (x, A) above (y, B), where x and y are, respectively, the outcomes you in fact achieve with A and B as your opportunity sets. After all, you may prefer A to B (for any of the reasons above), but if you would end up worse off with A than with B, say, because B is a much larger set and you quite consistently make poor decisions when faced with too many alternatives, this isn't irrelevant to our assessment of how instrumentally valuable to you A is as compared to B.

Vallentyne (2002) defends a somewhat similar view. He proposes to identify the value of opportunity sets with their expected level of outcome advantage. One opportunity set A is (instrumentally) better for you than another set B just in case your expected level of advantage, conditional on A being your opportunity set, is higher than your expected level of advantage, conditional on B being your opportunity set. Given a set of options to choose from, the outcome that an individual achieves is jointly determined by the individual's choices and the state of nature that obtains. Thus Vallentyne assumes that the chance of any outcome a is a function of the probabilities of the different states of nature and the probabilities of the different choices the individual might make.

An example might help illustrate. Suppose there is a 0.5 chance that Alice will choose to stay indoors reading a book and a 0.5 chance she will choose to go out for a walk, and that there is further a 0.5 chance that she will get wet if she goes out for a walk, and a 0.5 chance that she will remain dry if she goes out. Suppose Alice achieves a welfare level of 5 by going out and remaining dry, a welfare level of 2 by staying indoors, and a welfare level of 0 by going outside and getting drenched. Alice's expected level of welfare is $0.5 \cdot 2 + 0.25 \cdot 5 + 0.25 \cdot 0 = 2.25$. Thus, on this model, an option is valuable to you insofar as it improves your expected level of welfare; this is consistent with believing *both* that giving you additional options may be better for you even if you don't choose them or rank them first (say, because their mere availability increases the welfare you derive from other choices) *and* that giving you "too many" options may be bad for you (because it decreases the chance that you will choose the option that would make you best off or increases your decision-making costs).

Whatever the merits of these particular proposals, they all miss the importance of the diagnostic value of options. I noted earlier that one may have incomplete preferences, or that we may uncertain about our future preferences, and that for that reason one may wish to keep one's options open; but of course choice itself plays a large role in helping us to complete our preferences, and in reducing our uncertainty regarding the stability of our preferences. It's only by going skating that you can find out how you enjoy skating, and only then are you in a position to rank skating against skiing. Likewise, if you keep skiing regularly, and you find yourself enjoying it more and more every time, this should presumably increase your confidence that you will continue to prefer skiing over skating in the future. So although it's perfectly true that you might rationally want you and your friends to defer for another week your decision of whether to rent a ski lodge for the weekend or to rent a small chalet by a frozen lake, because by then (with enough introspection) you might make up your mind, you might also be rationally willing to pay *instead* for the additional option of going skating yourself for a few hours at the local arena, because by exercising this option, you might acquire just the information you need to make up your mind about the other two options.

It would be misleading, however, to say that the option to go skating for an afternoon is your top-ranked option: if the alternative is to save the money on buying a ski pass for the day, which you know you would enjoy, you may be unable to rank going to the arena against its alternative, and it may be rationally permissible for you to choose either, though your choice would be grounded in different reasons (to learn about your tastes, in one case, and to enjoy yourself, in the other). Nor is going to the arena necessarily what's best for you: if it turns out that you *really* wouldn't enjoy skating, but you just don't know it, then what would be best for you, from the sole point of view of welfare, would be to buy a ski pass. The value of the option to go to the arena is purely diagnostic in this case: all the good it does you is to teach you about yourself and to help form new judgements. Of course, like all knowledge, knowledge about yourself *may* instrumentally enhance your objective long-run expectation of welfare, by making you more likely to choose well later. But diagnostic value is not purely instrumental: life is a series of value judgements, and it is in itself part of good living to be a competent evaluator, one who knows their own values and tastes, is aware of the evaluatively salient features of their alternatives, and can more frequently and more confidently reach good judgements.

The diagnostic value of options will ultimately be baked into the all-things considered ranking of pairs of the form (x, A), but if, as the example above suggests, the diagnostic value of an option is distinct from the value it derives from being your top-ranked option, or its conduciveness to your welfare, then the diagnostic value of an option makes a distinct contribution to the overall value of the set, in the same way that the representative and symbolic value of one's options do. It may then be of interest to us to assess how diagnostically valuable different options are. And in general, since some sets of options might be richer in symbolically or representatively valuable options, while others may be richer in diagnostically valuable options, it may be of interest to us to assess how diagnostically valuable a set of options is, on the whole, so that we may be in a position to make informed trade-offs.

5.2 What is diagnostic value?

So far, I've only explored the concept of diagnostic value through examples. Now it's time to make this idea more precise. As I define it, the diagnostic value of an opportunity is the epistemic-*cum*-pragmatic value of the information generated by choosing it, in terms of making one a better judge of value, as measured by the (potential) changes one's preferences can undergo as a result of processing this information ("preference" here is synonymous with "subjective value assessment"). This concept of value raises four questions: (a) what kinds of preference changes are there? (b) how does one measure the diagnostic value of choice by these changes? (c) are all choice-induced preference changes diagnostically valuable, or only those which constitute learning of some kind? (d) what even constitutes preference learning? In this section I focus on (c) and (d), and will try in the next section to answer (a) and (b).

Regarding (c): only preference changes which constitute some kind of learning, or an improvement in one's faculties of judgement, are diagnostically valuable. Choices which merely cause preferences to change without providing a reason for the change are excluded. This is because diagnostic value is tied to the value of self-knowledge and of being a good judge of the value of one's options: a change in one's preferences must count as a cognitive achievement if it is to count as diagnostically valuable, and therefore must be the effect of an appropriate response to a learning event.

Thus, if I get drunk and my preferences change – say, I develop a strong desire to go shopping-cart surfing off my parents' roof – this does not constitute a cognitive achievement. I've not learned anything about myself (e.g. I have strong bones) or about shopping-cart surfing (e.g. it's not that risky) that would justify my taking a more favourable attitude to the prospect of launching myself in a metal crate from my parents' roof and onto the concrete. Rather, the alcohol has simply altered my brain chemistry. Likewise, suppose that upon learning that some desired outcome is less likely to happen, my preferences adapt and I come to desire it less. Here the preference change is responding to a learning event, but not in an appropriate fashion, as the information that is acquired doesn't seem to provide a reason for the change. As Bradley (2017, 200) puts it: "the fact that I am unlikely to win a competition is no reason to regard winning it as less valuable than winning at one in which my prospects are good." Indeed, Bradley (2017, p. 207) argues that rationality prohibits this sort of preference change.

So it's a necessary condition that a preference change constitute learning to be diagnostically valuable. But is it a sufficient condition? Not all change is good change, you might argue, even when it's the result of a learning experience: sometimes you get misleading evidence. Maybe Elizabeth saw a bad ballet performance, and would discover she actually likes ballet if she went to see good performances. Therefore, after seeing only one performance, her preferences have actually moved further away from the preferences she would have if she were much more informed. You might then think that a preference change should only count as diagnostically valuable if the change brings you closer to the preferences you would have if were maximally well-informed. Unfortunately, while this line has some persuasive force, it rests on the assumption that there is such a thing as the preferences you would have if you were maximally well-informed. However this is not so, because rationality frequently underdetermines how your preferences must change in response to a learning event (see section 7). Without a "true" preference relation to guide us, I conclude that our only criterion for judging whether a preference change is good is whether it was rationally mandated as a response to a learning experience. The point about misleading evidence is a red herring. As we will see later, Elizabeth should not form a confident value assessment about ballet based on a single viewing,

This brings us to (d): I take preference learning to be any kind of learning experience which informs our value judgements, and thereby rationally compels a change in one's preferences over actions and outcomes or (as we will see later) in one's confidence in those preferences. Of these, four may be distinguished. First, we may revise our preferences over possible courses of action upon learning something about the world or about ourselves which changes the expected utility of engaging in those courses of action – e.g, when you change your attitude to going to the beach after looking up the weather and learning it will rain (Bradley, 2017, p. 200). Second, we may revise our attitudes to certain outcomes upon learning something on which the desirability of the outcome depends – e.g. if I have already planned a vacation to Sydney, then learning that it will be sunny all week makes me look forward to the trip more than learning that it will rain (Bradley, 2017, p. 207). Third, we may change our attitudes towards some action or outcome by discovering through experience something about our own likes and dislikes – e.g., forming a preference to go skating after trying it once and discovering one enjoys it (Bradley, 2017, p. 201). Fourth, we may change our preferences by becoming aware of new possibilities or options – e.g. learning that, in addition to white, red, and rosé wine, there is also orange wine, may change your disposition to buy white.

The first two cases differ from the third and fourth in important ways, but all four are types of preference learning. In the first two cases, some propositional content is learned, which then alters our beliefs about the state of the world, and since what we desire to do or happen depends partly on what we know or believe will happen, this rationally leads us to change our preferences. In the third case, although some propositional content is surely learned (e.g. "I enjoy skating," or "wine is better with cheese", or even "ah, *this* is what wasabi tastes like!"), it isn't learning that content which prompts the preference change, but rather the positive or negative phenomenal valence of the experience itself. Still, you have learned *something* about the action or outcome you've experienced (its "what-it-is-likeness"), which can be represented by a proposition about how the outcome or action stands with respect to your tastes,[1] and in response to which there may be more-or-less appropriate ways of revising your preferences. For example, supposing you prefer to engage in activities you enjoy than ones you dislike, you should come to prefer skating over activities you dislike upon learning that you actually enjoy it, and likewise the intensity of your desire should for skating should change in the same direction as the change in your enjoyment of the activity.[2] In the fourth case there is also some propositional knowledge acquisition ("there is a fourth type of wine"), but in this case the learning experience doesn't merely modify our existing attitudes (e.g. decreasing my desire for white wine), it makes us aware of new ways the world might be and new options we might choose (I could buy orange wine instead of white). So-called "awareness growth" (see Karni and Vierø, 2013; Bradley (2017, chapter 12)) is a distinct kind of learning experience from the other three, and will require separate treatment later.

This ends our treatment of (c) and (d); we now know what diagnostic value is. Note that as I've characterized it, the diagnostic value of one's choices is different from their diagnostic *utility*, which is how Bodner and Prelec (1997, 2002) refer to the utility of the information which one's choices reveal about our own traits

[1] Representing preference changes in sentential form is a common move in the literature on preference change. See e.g. Hansson (1995), Bentehm and Liu (2007)

[2] This remark connects naturally with arguments in defense of the desire-as-belief thesis, which states that a rational person desires some proposition to be true just to the extent that she believes or expects it to be good. See Broome (1991a), Bradley and List (2009), and Bradley and Stefánsson (2010) for discussion and defense.

and dispositions.[3] To illustrate, they suggest that someone who chooses to go for their daily jog in spite of rain may derive diagnostic utility from this choice, because jogging in spite of the rain is taken as evidence that they are steadfast and strong-willed, which is a desirable trait to have. To be sure, learning about one's traits can help us form new preferences (e.g. learning that I don't have an addictive personality makes it more attractive to use drugs recreationally every now and then). But it hardly needs emphasizing that the value of news conveyed by choice in terms of making us better evaluators cannot be reduced to how much we desire that news: I may not *like* discovering that I don't enjoy skating – what kind of Canadian does that make me? – but I am now in a better position to judge what to do next weekend. That said, the extent to which we care about the news our choices convey may affect the latter's diagnostic value. And this brings us to (a) and (b): what sorts of preference changes can result from preference learning, and how does one measure diagnostic value by these changes?

5.3 Dimensions of diagnostic value

If what makes an option diagnostically valuable is that the information which choosing it would generate can improve your evaluative capacities, provided you revise your preferences appropriately, then presumably an option is more diagnostically valuable the greater the size of this improvement – that is, the greater the change in your preference structure rationality mandates as a response to your learning experience. Different changes are possible, though, making diagnostic value multidimensional in nature. "Simple" preference change, in the sense of forming new preferences or reversing one's existing preferences, is different from awareness growth. And both are different still from changes in *confidence* in one's preferences, which may also result from a preference learning experience, or so I will argue, even if the experience does not otherwise alter those preferences. Moreover, as I will argue later, the diagnostic value of a choice also depends on how much we care about the information it generates or the improvement it enables. Thus, diagnostic value has four dimensions, and in the following sections I will look at each in turn, starting with simple preference change, followed by awareness growth, confidence change, and how much we care about the information generated by our choices.

To make this conversation precise, however, we ought to start with a model of an agent's preference and belief structure, so that we can model (and measure)

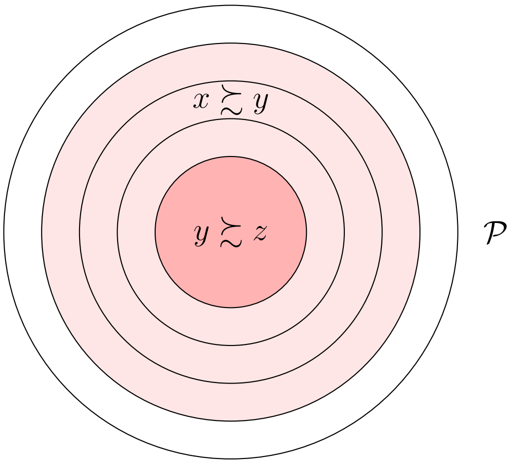
[3] This concept is closely related to how observation is defined in the study of causal networks (Pearl, 2000). See Bovens (2013) for discussion.

changes to this structure. I will adopt the framework developed by Hill (2013); this is a non-standard representation of individual attitudes, but it is sufficiently general to model all the changes of interest to us under a single representation. As ever, we start with a set $X = \{x, y, ...\}$ of mutually exclusive possible alternatives, and a set Z of nonempty subsets $A, B, ...$ of X, called opportunity sets. We define a set $\mathcal{P}$ of possible preference orderings over X; each element of $\mathcal{P}$ is denoted $\succsim_1, \succsim_2$, and so on, and subsets of $\mathcal{P}$ are denoted by the generic terms $\mathcal{R}_1, \mathcal{R}_2$, etc. Each $\succsim_i \in \mathcal{P}$ is assumed complete and transitive, and so represents a fully determine preference ordering an individual might have.

The distinctive feature of Hill's framework is that now, instead of using just a single preference relation to represent the individual's preferences, we use a *nested family* of preference relations – that is, a set of subsets of $\mathcal{P}$ such that, for any two distinct subsets, one is contained in the other. Let Ξ denote such a nested family of subsets of $\mathcal{P}$. If there is a set of weak orderings $\mathcal{R} \in \Xi$ such that for every $\succsim_i \in \mathcal{R}$, $x \succsim_i y$, then the individual weakly prefers x to y. The individual's weak preference relation is denoted $\succsim$, and it can be recovered by looking at the smallest $\mathcal{R} \in \Xi$ (i.e., the smallest set of preference orders such that every other distinct set of preference orders belonging to Ξ contains it): if $\mathcal{R}$ is a singleton $\{\succsim_i\}$, then $\succsim = \succsim_i$, and $\succsim$ is accordingly complete. Otherwise, $\succsim$ is the intersection of all the $\succsim_i \in \mathcal{R}$, and it is incomplete. And intuitively, $\succsim$ denotes the individual's best guess as to how any two elements in X compare. Thus when the smallest $\mathcal{R} \in \Xi$ is a singleton, the individual always *has* a best guess as to how two options compare, but when $\mathcal{R}$ is not a singleton, then even the individual's best attempt at ordering X leaves some pairs of elements unranked – i.e., the individual can't do better than suspend judgement in some cases.

Of course, even if the individual can reach a determinate value assessment, they may not be confident in the assessment they've reached. So even if $x \succsim y$ and $y \succsim z$, the individual may be more confident in the first value assessment than in the second, and thus may be willing to rest more weight on it in high-stakes situations. Hill proposes that degrees of confidence in the assessment that $x \succsim y$ can be captured by the largest set $\mathcal{R}_i \in \Xi$ such that $x \succsim_i y$ for all $\succsim_i \in \mathcal{R}_i$. The individual is more confident that $x \succsim y$ than that $y \succsim z$ if the smallest set $\mathcal{R}_i \in \Xi$ such that $x \succsim_i y$ for all $\succsim_i \in \mathcal{R}_i$ includes but is not included by the smallest set $\mathcal{R}_j \in \Xi$ such that $y \succsim_i z$ for all $\succsim_i \in \mathcal{R}_j$. We say that a preference order $\succsim_1$ is more *plausible* (i.e., the individual thinks, in light of their experiences, that it's a better "best guess") than another preference order $\succsim_2$ if the smallest $\mathcal{R}_1 \in \Xi$ containing $\succsim_2$ is a superset of the smallest $\mathcal{R}_2 \in \Xi$ containing $\succsim_1$. For any nested family of preference orders Ξ, there corresponds a unique "implausibility order," noted $\trianglelefteq$, on $\mathcal{P}$, which ranks preference orders by their implausibility. Plausibility is closely tied to confidence: an individual is more confident in some value assessment the higher degree of implausibility up to which all preference orderings agree on this assessment. This terminology allows us to define the notion of a preference relation *at a level of confidence*; this is identified with the intersection of all the $\succsim_i$ up to a certain degree of implausibility.

A graphical representation may help. In the figure below, every point on the plane represents a possible weak ordering. Each colored disk is a set of preference orders belonging to Ξ, so if a value assessment holds in some disk that means it holds at every point in that disk. The darker disk represents the individual's "best guess" as to how the elements in X compare; this disk would be a point if the individual's best guess were a complete preference order. If one value assessment holds in a bigger disk than another, this means the individual is more confident in it. Note that this diagram also makes clear that there are a lowest and a highest point of confidence: the lowest point is reached whenever some value assessment only holds in the darkest disk (i.e., every $\succsim_i$ in the smallest $\mathcal{R} \in \Xi$ agree on it, but not in any superset of $\mathcal{R}$), and the highest point is reached whenever a value assessment holds at all points in the largest coloured disk (i.e., when all $\succsim_i \in \cup_{\mathcal{R}_i \in \Xi}$ agree on the assessment). Confidence being inversely related to plausibility, this implies that there is a highest and a lowest level of plausibility too.



Although Hill developed this framework to model individual preferences, Bradley (2017, chapter 14.3) shows it can also be used to model individual beliefs. In this case, instead of a set X of possible alternatives, we have a boolean algebra $\Omega = \langle \mathcal{X}, \models \rangle$ – i.e., a pair consisting of a set $\mathcal{X}$ of propositions which is closed under negation, conjunction, and disjunction, and an implication relation $\models$ on that set of propositions which ensures that the distribution of truth values over $\mathcal{X}$ obeys classical logic. We can think of Ω as the set of all possibilities one can have attitudes about. X and Ω are connected, however, in that X can be interpreted as a partition of Ω, namely, that consisting of the set of propositions that the individual could make true. Instead of a nested family of sets of preference relations, we have a nested family of sets π_1, π_2... of probability distributions $p_1, p_2, ...$ over Ω. Each p_i represents one possible complete and coherent set of beliefs about the world. The p_i are ordered as before in an implausibility order, and the individual's best guess as to how the world is corresponds to the smallest π in the nested family.

5.4 Formation, Restriction, Revision

The presentation of Hill's framework being now complete, we can look at the relation between diagnostic value and simple preference change. There are three kinds of simple preference changes: preference formation, in which we reach a definite value assessment between two objects about which we previously reserved judgement; preference restriction, where we withdraw some value assessment regarding two objects and now suspend judgement; and preference revision, where we reverse our value assessment regarding two objects.[4]

The cases I used to introduce the concept of diagnostic value were ones of formation: individuals start out not knowing how to rank two possible options, then, after choosing to perform some act from among a list of alternatives, they learn something about their own likes and dislikes and make up their mind, filling in some of the gaps of their preference ordering. Thus, as a result of choosing x from the opportunity set A, the individual's preference relation $\succsim$ gets extended to an updated preference relation $\succsim_{x,A}$, which is a strict super-relation of the prior relation $\succsim$. Note that I index the updated relation not just to the option which was selected but also to the opportunity set it was selected from, because the information generated by choosing x may not always be the same depending on what alternatives were available: choosing to donate blood when you could choose not to may teach you something about the value of altruism which you could not learn if donating blood were your only option, since it is part of what makes some personal sacrifice altruistic that it was up to you whether to make this sacrifice or not.

Acquiring new preferences is valuable to decision-makers because it releases them from indecision and caprice, but sometimes preference restriction is the appropriate response to what one learns. Imagine a committed activist and volunteer

[4] In the dynamic logic of preference change revision is not distinguished from formation, as both are modelled by the addition of some sentence to a set, which must then be further altered by dropping any sentences that are inconsistent with the addition (see Hansson, 1995; Bentehm and Liu, 2007). The intuitive difference between formation and revision gets captured by the fact that in the case of formation, the addition will not contradict any existing sentence in the set.

who is, at first, confident in her belief that the best plan of life for her is to join an NGO like Oxfam or Amnesty, but who, after watching a remarkably chilling televised report on the brutal military occupation of Western Sahara, becomes deeply impressed by the work of the journalists behind the report, and begins to wonder if she shouldn't instead aspire to become a war corespondent. This change in our activist's preferences should be regarded as a diagnostic improvement: making up one's mind about an issue can be a valuable change, but so can realizing that one had made up one's mind too hastily. By re-examining one's priorities, one may avoid making rash or hasty decisions. In these cases, the updated preference relation will be a strict subrelation of the prior preference relation.

Now consider the case of preference revision. Suppose that I currently prefer playing online chess to reading poetry. Poetry only mildly entertains me, but I've been a serious chess player for some time – I even pay for a subscription to chess.com. But one day, on a whim, I read the opening lines of *The Lovesong of J. Alfred Prufrock* and, mesmerized, continue reading it to the end, then read it over and over again. This experience has taught me something very important about the value of poetry, and this will lead my desire for online chess versus poetry to weaken considerably. If the price of my premium subscription to chess.com increases, I may be just as happy to downgrade my membership, give up playing so much chess, and read more poetry instead. In this case the diagnostic value of the option to read *The Lovesong of J. Alfred Prufrock* is that choosing it forced me to reverse my preferences over certain combinations of time spent playing chess and time spent reading poems. As I continue reading more great poems my preferences over more combinations will shift until I generally prefer poetry to chess.

Note that although I've only used examples where one learns about one's tastes in discussing preference change, any of the first three preference learning experiences canvassed in section 3 could lead to preference formation, restriction, or revision. In the case of formation, for instance, the incompleteness in your value assessments could be due either to the fact that you don't know how two objects stand with respect to your tastes, or to the fact that their comparative desirability depends on the state of the world, which you are uncertain about.

I won't go into the details of how rationality requires you to update your preferences or degrees of desire in response to new information (though for illuminating discussion, see Hansson 1995; Dietrich and List 2012; and Bradley 2017, ch. 10.6), because the task before us now is to measure the diagnostic value of an option along the dimension of simple preference change, a dimension noted **p**, and all I need to define this measure is the assumption that rationality does sometimes require us to revise our preferences, which is uncontroversial. Intuitively, the bigger the change rationality requires in your preferences as a result of some learning experience, the more diagnostically valuable the choice leading to this experience was. If going to the arena *only* makes me realize that I prefer skiing to skating, but going to see *Manon* helps me realize that I prefer to see an Opera, a play, a concert, or a musical to going to see the ballet, then it seems that going to the ballet is in one respect a more diagnostically valuable choice to me. Diagnostic value is increasing in the size of the improvement to one's capacities of judgement choice enables.

This idea can be made precise. Let $Z(\succsim, \succsim_{x,A}) \subseteq X \times X$ denote the set of all pairs $x, y \in X$ on whose ranking $\succsim$ and $\succsim_{x,A}$ disagree. So, if your preferences update towards completion, there will be some pairs which $\succsim$ refuses to rank but which $\succsim_{x,A}$ ranks, and vice-versa if your preferences update towards restriction, and if your preferences reverse there will be pairs which $\succsim$ ranks in reverse order to $\succsim_{x,A}$; all these different pairs are elements of $Z(\succsim, \succsim_{x,A})$. We now say that x is as least as diagnostically valuable as y along $\mathbf{p}$ just in case $|Z(\succsim, \succsim_{x,A})| \geq |Z(\succsim, \succsim_{y,B}$ $)|$. Or, if X is infinitely large, but measurable, the measure of each $Z(\succsim, \succsim_{x,A})$ is substituted for its cardinality. This rule induces a complete ranking, and has the attractive implication that some option has no diagnostic value along dimension $\mathbf{p}$ if and only if your value assessments would rationally remain unchanged from choosing it.

5.5 Confidence change

John Maynard Keynes (1973/1921, p. 73) introduced a distinction between what he called the weight of evidence and what we might call the balance of evidence. The balance of evidence is the degree to which one's evidence supports a particular proposition, i.e., how high a probability assignment in some proposition it justifies. The weight of evidence is how substantial the evidential basis for some probability assignment. To get an intuitive handle on the distinction, consider an example from Popper (1959/1934): suppose I show you a coin and ask you to bet on its landing heads heads; in your ignorance of whether it's a fair coin, you might reasonably apply some version of the principle of indifference and assign equal probabilities to its landing heads as to its landing tails. If I now flip it a thousand times, and it lands heads exactly 500 times, then you will continue, presumably, to assign equal probabilities to its landing heads as to its landing tails, but now your judgement is made on a much sounder basis. The balance of evidence is the same, but the weight of evidence has greatly increased.

What goes for beliefs goes for preferences. Suppose I've loved the Beatles since I was a child, and I just think they're the best act in rock history. Over the years, I've listened to an ever widening repertoire of rock music, some of which I have loved greatly, some of which left me cold, but every new experience has only served to ingrain my long-held belief that the Beatles are the best, that their music is better than the rest. My preference for the music of the Beatles is now very strongly confirmed by the evidence I have accumulated over my years of listening to music, and I can anticipate that it will remain stable over the years to come. Notice that this need not imply that the relative intensity of my preference for the Beatles over other bands about which I had prior opinions has changed: Pink Floyd may always have been a close second. In fact, as my tastes expand, and I become more opinionated, I may spend relatively less of my time listening to the Beatles, and relatively more listening to the hundreds of other artists that I've discovered. Still, I continue to prefer the Beatles to any other act, *Abbey Road* remains my favourite album, and *Eleanor Rigby* my favourite song ever composed in the dorian mode. In this case, the balance of evidence over the course of my life has always supported the value judgement that The Beatles dominate all other acts, but the weight of the evidence underpinning this judgment has substantially increased.

The distinction between the balance of evidence and the weight of evidence maps on to the distinction between the individual's preference relation and their implausibility order. Preference orders which fail to rank the Beatles first become less and less plausible as I accumulate more musical experiences which are inconsistent with those preference orders, and so I become more confident in my childhood assessment that the Beatles are the best. Naturally, as I undergo more preference learning experiences, my confidence in some value assessments may drop, but in that case it's my confidence in other value-assessments that will rise. All increases in the weight of the evidence underpinning my value assessments will therefore be recorded as changes in my implausibility order.

The question for us is whether changes to an implausibility order induced by preference learning should count as diagnostically valuable. I think so. It's perfectly fine to make decisions on the basis of somewhat lightly-reached assessments of value and probability when the worst that can happen if you've reached the wrong assessments isn't so bad. But when the price for misjudging the best course of action is steep, and we really don't want to get it wrong, greater caution is required in reaching one's assessments – i.e., greater confidence in the assessments – and it's quite rational to punt the decision to later, accepting a sure loss to buy oneself the time to gather more evidence. Indeed, caution of this sort is part of the motivation for the imprecise Bayesian approach (see e.g. Levi, 1990; Gärdenfors and Sahlin (1982); Gilboa and Schmeidler, 1982; Binmore, 2008). Having a weightier evidential basis for our value assessments makes us more competent judges of value, because it means we remain capable of rendering determinate value judgements across a wider range of circumstances, and particularly in higher-stakes situations.

Moreover, embedding confidence in the concept of diagnostic value links the latter quite naturally with Pattanaik and Xu's remark about our uncertainty regarding our future preferences. Presumably, if I am uncertain about the diachronic stability of my preferences, this will be reflected in the fact that my current preferences lack much confidence. And conversely, if I am quite confident in my preferences, then I should be more confident that they will remain stable over time, and this puts me in a better position to evaluate how much I should be willing to pay to keep my future options open. Thus there are all sorts of decision where evaluative competence requires having a weightier base of evidence on which to rest one's judgements.

As before, I presume that the diagnostic value of an option is increasing in the increase of the weight of our evidence – and therefore in the size of the change to our implausibility order – that choosing it would provide. Since $\trianglelefteq$ is a weak ordering, we can define as before the set $Z(\trianglelefteq, \trianglelefteq_{x,A}) \subseteq \mathcal{P} \times \mathcal{P}$ as the set of all pairs $\succsim_i, \succsim_j \in \mathcal{P}$ which $\trianglelefteq$ and $\trianglelefteq_{x,A}$ disagree about how to rank. We now say that x is at least as diagnostically valuable as y along the dimension $\mathbf{c}$ of confidence if and only if $|Z(\trianglelefteq, \trianglelefteq_{x,A})| \geq |Z(\trianglelefteq, \trianglelefteq_{y,A})|$. The measure of the sets is substituted for their cardinality if the sets are infinite but measurable. As before, this measure has a natural zero point: some option has no diagnostic value along dimension $\mathbf{c}$ if and only if your implausibility assessments would rationally remain unchanged from choosing it. This ranking is also complete.

Note that any change in the individual's preference relation implies a change in their implausibility order,[5] so an option that is diagnostically valuable along $\mathbf{p}$ must also be diagnostically valuable along $\mathbf{c}$. Still, having both measures is useful, as you might think that changes to your implausibility order are more diagnostically valuable when they also imply changes to your "best guess" at the right value assessments, and therefore that the former changes should receive greater weight in the measure of overall diagnostic value. Your best guess is special, as compared with your best guesses at higher levels of confidence, because it's your *best* guess, your fullest attempt at a complete and coherent evaluation of the world. And of

[5] This is easy to show: $\succsim$ is defined as the intersection of all the $\succsim_i$ belonging to the smallest set $\mathcal{R} \in \Xi$, thus $\succsim \neq \succsim_{x,A}$ iff the smallest $\mathcal{R} \in \Xi$ is different from the smallest $\mathcal{R}' \in \Xi_{x,A}$. But if $\Xi = \Xi_{x,A}$ then $\mathcal{R} = \mathcal{R}'$.

course, it's possible for one option to be more valuable along **c** than another yet less valuable along **p**,[6] so it's far from redundant to retain both orders.

5.6 Awareness growth

A final respect in which choice may inform our value judgements is by making us aware that there are more ways of intervening upon the world and more ways the world could be than we had previously countenanced. That is, we could learn that the set of possible options is more expansive than we thought it was – e.g. I learn from soliciting the advice of a sommelier that in addition to bubbly, white, red, and rosé wine, I could also order orange wine, which is not a type of wine I previously knew existed. Or we could learn that there are more possible states of the world than we had realized, on which the desirability of other states we were previously aware of may depend, and which may yield consequences we had not considered, depending on what course of action we pursue – e.g. I learn that in addition to rain and sunshine, the weather in Canberra may also turn to violent hail and asphyxiating smoke haze.

Strictly speaking, there are two varieties of awareness growth: expansion, where the new possibilities one becomes aware of are disjoint from any of the possibilities we were previously aware of, and refinement, where they are not. So in the case of options, learning that one can also order orange wine in addition to bubbly, white, red, and rosé wine is a case of expansion: to buy orange wine is ipso facto not to buy any of the other four kinds of wine. In contrast, if I start being unable to make fine distinctions between types of white wine, but then, through experimentation, learn that there is an appreciable difference between Sauvignon Blanc, Pinot Grigio, Viognier, and Chardonnay, then we have a case of refinement; I go from having a coarser-grained representation of the space of possible options (e.g.{red, white,...}) to a finer-grained one (e.g.{red, Sauvignon blanc, ...}). One way of putting this is that with expansion, I learn that more options are possible than I thought, and with refinement, I learn that the space of possible options is more diverse than I thought.

[6] Suppose the smallest set in Ξ contains only two preference orders which disagree with each other about everything; your preference relation is therefore empty. Now suppose choosing x would teach something which makes one of those two preference orders less plausible, so your updated preference o is identical with the remaining order. This implies a maximal change to your preferences, but a rather small change to your implausibility order, since your updated implausibility order and your prior implausibility order only disagree about the ordering of two preference orders. In contrast, suppose choosing y would not affect your judgements as to the plausibility of the preference orders you find most plausible, but would change your judgements about many of the less plausible preference orders. Then choosing y implies greater changes to your implausibility order than choosing x, but no changes to your preferences.

In the case of the possible states of the world, learning that hail and smoke haze are additional possible weather states is a case of expansion, since each type of weather event is mutually exclusive with every other. In contrast, suppose the possible weather states you are aware of for tomorrow (rain, sunshine, hail, smoke) exhaust the space of possibilities, but you also believe that whether it will be hot or cold also affects the desirability of your various options; temperature can vary independently of weather state, so you must assess the likelihood and desirability of your doing so-and-so when it is rainy and cold, rainy and warm, etc. If now you inform yourself and learn that humidity may also vary independently of temperature and weather, and the humidity level affects the desirability of your courses of action, then you must now refine your attitudes to consider the consequences of doing so-and-so when it is cold and sunny and humid, cold and sunny and dry...

Some decision theorists try to treat all cases of expansion as cases of refinement by defining a "catchall" proposition (Shimony 1970), which stands for all possibilities one hasn't considered.[7] This may be reasonable in some cases, as when attempts to diagnose gaps in knowledge predict unanticipated possibilities but cannot identify them (Loch, Solt, and Bailey, 2007), because in this case it may be possible to make (imprecise) probability estimates by looking at the size of the gap in knowledge. Barring those cases, though, a catchall is an unreasonable strategy, because, as Bradley (2017, p. 255) argues: "given that we don't know anything about the prospects that we are potentially unaware of, (...) what probability should we assign to the catch-all prospect? But if we recognise the lack of such a basis then we will be driven towards a state of radical non-opinionation, which does not seem to be much of an improvement over that of conscious unawareness."[8]

There are differences in how it is rational to adjust our attitudes in response to expansion and refinement. In the case of refinement, it is in principle possible to extend one's attitudes to the new finer-grained possibilities being considered without modifying one's attitudes to the old coarse-grained possibilities: learning

[7] In Bayesian confirmation theory, awareness growth is related to the problem of new hypotheses (Glymour 1980). In that context the catchall stands for all hypotheses not considered.

[8] There is a growing project in the field of risk management around quantifying the "unknown unknowns," i.e., putting a probability on the catchall. But in fact the methodology for doing so basically involves trying to identify the unknown unknowns, thereby converting them into known unknowns, i.g. anticipated contingencies whose probabilities we can try to estimate. And indeed Kim (2012) argues that in the case of unanticipated disasters like hurricane Katrina or the Fukushima nuclear accident, many of the contingencies which were supposedly unidentifiable in advance were in fact identifiable. So really, quantifying the unknown unknowns doesn't involve estimating the proabability of a catchall, but attempting to grow one's awareness and then estimating the probabilities of the possibilites one is aware of.

that white wine comes in multifarious varieties need not change my opinion that white wine is better than red. In the case of expansion, however, your attitudes to the old possibilities must be modified to accommodate your awareness growth: learning of some new disjoint possibility must reduce the probability you assign to at least one of the previously countenanced possibilities, and this in turn will affect the expected utility of your available options, and likewise learning of some new possible option will necessarily change how much you desire to pursue your other options whenever you aren't indifferent between the new option and any of the old ones. In either case, rationality underdetermines how one must distribute one's credences: if my awareness grows by expansion, I can award either a huge or a tiny lump of probability to the new possibility, and rationality has nothing to say about which is best (in the same was that rationality imposes no constraints on how we set our priors).

Once again, I won't delve into what the rational constraints are on how one is permitted to revise one's attitudes following awareness growth, but will simply emphasize what matters for our purposes, which is that the process of attitude revision can be envisioned as a two-step process. First we grow our awareness of the possibilities, then we form new attitudes to the possibilities we are now aware of. This second step was discussed earlier, so the question for us is whether the first step constitutes an improvement in one's capacities of judgement, and therefore whether preference learning experiences which result in awareness growth are diagnostically valuable for that reason. Clearly so: simply being aware of potentially relevant new possibilities better equips us to make wise decisions, even if we cannot yet form determinate evaluative or probabilistic judgements about these possibilities. If nothing else, it will alert us to the need for caution and informing ourselves.

Measuring the diagnostic value of an option along the dimension of awareness growth is fairly straightforward. Diagnostic value being monotonic in the size of the improvement to one's capacities, and greater awareness being a greater improvement, it follows that choice is more diagnostically valuable the more possibilities it makes us aware of. If $\Omega = \langle \mathcal{X}, \models \rangle$ is the background boolean algebra, let $\Omega_{\mathcal{A}} = \langle \mathcal{A}, \models \rangle$ with $\mathcal{A} \subseteq \mathcal{X}$ be the set of all possibilities of which the agent is aware prior to the moment of choice. Finally, let $\Omega_{\mathcal{A},x,A}$ denote the set of all possibilities the agent is aware of after having chosen x from A, given that $\Omega_{\mathcal{A}}$ was their prior awareness state. Then, we say that $x \in A$ is more diagnostically valuable than $y \in B$ along the dimension $\mathbf{a}$ of awareness growth if and only if $|\Omega_{\mathcal{A},x,A}| \geq |\Omega_{\mathcal{A},y,B}|$. If the sets of possibilities are infinite, but measureable, their measures are substituted for their cardinality. This ranking is complete.

5.7 Information we care about

Diagnostic value is not purely a function of what we learn, but also of our attitudes to what we learn. As Bodner and Prelec emphasize, learning experiences which inform us of desired traits and dispositions are more valuable to us than ones which does not. Moreover, preference learning experiences can be more-or-less desirable if the news allows me to form new opinions about things I care more-or-less about. If I enjoy wine, then it may matter to me greatly that I be able to make fine distinctions between types of wine, since this may put me in a much better position to make good decisions. In contrast, if I hate beer, finding it unbearably bitter, then I may be uninterested in trying out every different kind of beer out there. Thus, if choosing x will enable me to refine my attitudes over wine, and choosing y will enable me to refine my attitudes over beer, then, *ceteris paribus*, x should be considered more diagnostically valuable: whether or not x teaches me more about my attitudes than y, what it teaches me is worth more to me.

Likewise, if I am, in general, a great enthusiast of stage art, then the fact that I cannot rank ballet against any other form of stage art may cause me some distress; for all I know, I'm missing out on truly unique and wonderful experiences, thus it may matter to me to acquire the information necessary to make up my mind about these things. In contrast, if I am generally indifferent to soap operas and find myself unable to rank watching an hour of *All My Children* against an hour of *General Hospital*, I may be in no particular rush to fill that lacuna in my preference ordering. Thus, I prefer to learn such information as will help complete my preferences over stage art than as will help me complete my preferences over soap operas, and accordingly an option which would yield diagnostic news about the former, if chosen, should be seen as more diagnostically valuable than an option which would yield diagnostic news about the latter, if chosen. One could go on to consider similar examples regarding the other dimensions of diagnostic value. The point is clear enough: we may have preferences over news contents and over news types, and this itself should affect our judgment of how diagnostically valuable any given option is.

It might be objected that the extent to which we care about diagnostic learning events is not really a distinct dimension of diagnostic value. Sure, we may be more interested in improving our wine-related preferences than our beer-related preferences, but all this really shows is that the measure I first proposed in section 5 was inadequate, and that diagnostic value with respect to $\mathbf{p}$ should be a function not just of the absolute number of dyadic preference changes that occur following a learning event, but also how much you care about these changes. Ditto for the other dimensions. This interpretation could be motivated for a concern with conceptual unity: instead of three learning-related dimensions of diagnostic value and one preference-related dimension, we just end up with three learning-related dimensions. In fact, I wouldn't mind this move. However I'm not certain that the formal difference amounts to a very substantive one, since either way overall diagnostic value will end up depending on learning-related stuff and on preference-related stuff. In the meantime, I believe it clarifies my exposition to treat all four issues separately.

This is because the issue of how the diagnostic value of some option depends on our attitudes is a complicated matter: we must confront the problem that some people don't want to be better informed. Consider the case of a homophobic religious zealot, who, on the insistence of his liberal niece, goes to see *The Miseducation of Cameron Post*, and is deeply moved. As a result, he grows less confident in his homophobic value assessments and more confident in the "heretical" value assessments of his niece. Our zealot might not welcome this change, since it deeply challenges his entire world-view and sense of identity. He might therefore rationally, albeit regrettably, refuse to probe any further his homophobic value judgements through learning and experience. After all, from his present evaluative vantage point, he is like Ulysses anticipating the unwelcome future corruption of his preferences. This suggests that any preference learning experience which would further disconfirm his homophobic preferences and increase his confidence instead in "heretical" value assessments must be of very low (perhaps even negative!) diagnostic value, with respect to the dimension **d** of the desirability of preference learning events.

And yet if our zealot did undergo such a preference learning event (e.g. he allows himself to be roped into going to the gay pride parade with his niece), then, *ex post*, he would presumably be thankful for having made that choice, since now his preferences and his beliefs have changed in a way he cannot but regard as an improvement, since they changed as a result of receiving information in light of which he himself regards it as appropriate to revise his judgements. And he would therefore be inclined to regard his old beliefs and preferences as mistaken.

In situations like these, when our *ex ante* and *ex post* attitudes to the importance of a bundle of diagnostic news differ, precisely because acquiring the information in question changes our attitudes, it's not immediately obvious which set of attitudes is authoritative in determining the diagnostic value of options along **d**. On the one hand, it is on the basis of our prior preferences and beliefs that we make choices; when we choose x over y because of the value we attach to the diagnostic news conveyed by x, the preference we thus reveal is our *current* preference for x over y. In general, the only rankings over pieces of diagnostic news which we can empirically elicit from individuals are those induced by their prior preferences.[9] Moreover, it is presumably supposed to give us a reason to choose x that x is diagnostically valuable, but the Ulysses point stands: surely it gives me *no* reason to choose x that it will (as I presently see it) corrupt my preferences. This is very much related to the problem of how rational choice is even possible among actions when some of these would result in "transformative experiences," which change the values held prior to the choice (Paul, 2014; Callard, 2018).

On the other hand, if our prior and posterior preferences over bundles of diagnostic news disagree, then our prior preferences are formed in error, given that they are formed in the absence of precisely that information which we would later deem it necessary to have for a proper appreciation of the choice. Why not then simply defer to our better informed posterior preferences over bundles of diagnostic news? After all, when individuals disagree about what to believe, but one is known to be better informed than the other, and they are otherwise equally good at making inferences from available evidence, it's generally thought the less-well informed individual ought simply to accept the judgments of the better informed individual (this story wouldn't satisfy Ulysses, but then in Ulysses's case it's a bit unclear whether the sirens' song is a reason for his preference to drown or merely a cause of it, in the same way that snorting cocaine is a cause of addictive desires, but not a reason for them: many addicts report not enjoying their addiction and judging they should stop). Thus individuals' posterior preferences seem to be better indicators of the individual's "true" valuation of bundles of diagnostic news.

Now, different choices are bound to induce different preference changes, so we need to decide which of the possible posterior preference relations is the privileged indicator of an agent's "true" valuations; presumably it is the preference relation they would have if they were maximally well informed, or the preference relation that would be obtained by updating on every bundle of diagnostic news they could ever obtain. The problem, of course, is that such ideal circumstances never obtain, and so it is impossible to elicit empirically any individual's "true" valuations, and anyway there is no unique preference relation that the individual is guaranteed to form if they become maximally well-informed. So at best, we can make more-or-less well informed guesses about how individuals' preferences would likely change if they made this or that choice. We thus face a hard trade-off between empirical soundness and accuracy, on the one hand, and epistemological justification, on the

[9] The empirical elicitation of preferences may be done either by observing actual choice behavior, or by asking individuals to rank bundles of diagnostic news and gambles over bundles of diagnostic news, as in the Von-Neuman Morgenstern method.

other, in our rankings in deciding whether to take prior or posterior preferences as authoritative.

In any case, regardless of which set of preferences we take to be authoritative, it is simple to rank options along **d**. Let $E_{x,A}$ denote the preference learning experience undergone by choosing x from A. Quite simply, $x \in A$ is at least as diagnostically valuable along **d** as $y \in B$ if and only if the individual prefers to undergo learning experience $E_{x,A}$ than learning experience $E_{y,B}$. Notice that under this ranking rule, there is also a very natural zero-point. If some option x conveys no information to you (the "learning" event involves no learning), then it should also have no diagnostic value along **d**. As Jeffrey (1965, chapter 5.7) argues, you can't well desire to learn what you already know: news is only good or bad news compared with no news at all. If the individuals' preferences over learning events are incomplete, the ranking of X along **d** will be incomplete, but provided the individual satisfies all the remaining axioms of rational choice theory besides completeness, the diagnostic value of options along **d** will nonetheless be measurable by a family of utility functions. An option will be considered at least as diagnostically valuable as another along **d** just in case every utility function in the family assigns it at least as high utility. Note, though, that an individual's preferences over learning events might be complete even if their preferences over all possibilities are not.

5.8 Overall diagnostic value

Having discussed the measure of diagnostic value along each of its dimensions, we've now finished our treatment of (a) and (b). All that's left to discuss is how one might condense the disparate measures I've defined into a single measure of the overall diagnostic value of an opportunity set. There are two ways in which we might do this: we might either start by extending our dimension-specific measures of option diagnostic value to dimension-specific measures of opportunity set diagnostic value, then aggregating these dimension-specific measures into an overall measure of diagnostic value. Or we might start by aggregating the dimension-specific measures of option diagnostic value into a unique measure of overall option diagnostic value, then extending this aggregate measure to a measure of opportunity set value. The order of aggregation often matters, so it's not obvious in this case that these two ways of doing things will always be equivalent. Luckily for the approach I develop here, the order doesn't matter, so let us start by aggregating dimensions across options.

The following restrictions on the concept of overall diagnostic value seem reasonable. First, overall diagnostic value should be increasing in each dimension of diagnostic value: all else being equal, an option that is more diagnostically valuable along some dimension is more diagnostically valuable overall. Second, diagnostic value should be separable across dimensions, meaning that if two options are unequally diagnostically valuable across some dimensions, but equally valuable along remaining dimensions, then the ranking between them is invariant to just *how* valuable both are along the dimension along which they are equally valuable. This is just a way of saying that diagnostic value is additive, that an increase in diagnostic value along some dimension always counts for the same regardless of how diagnostically valuable some option already is along the other dimensions. Third, diagnostic value varies continuously along each of its dimensions. Thus no dimension lexically dominates the others, and trade-offs between dimensions of value are possible, such that a decrease in value along one dimension can be compensated by a large enough increase along any other dimension.

The first condition is a no-brainer. The second and third are not self-evident, but both seem plausible in the present context. What would a violation of additivity look like? Suppose we represent the diagnostic value of some option (or set) by a vector of four numbers. A non-separable ranking rule might rank (10,7,0,0) over (8,8,0,0) but rank (8,8,8,8) over (10,7,8,8) due to the perfect equality in the size of the diagnostic improvements along each dimension of the third vector. In contrast, a separable ranking rule that ranks (10,8,0,0) over (9,9,0,0) must rank (10,7,8,8) over (8,8,8,8). Now, nothing in our discussion to this point suggests that there might be interaction effects between dimensions of diagnostic value, and it's hard to see what rationale there could be for thinking that an option should be especially valuable for generating equally sized improvements along each dimension of diagnostic value. The distribution of diagnostic values across an option is not like the distribution of welfare or opportunities for welfare across society: there is no story to tell here about fairness, non-domination, or the special concern we owe to the worse-off.

As for the third condition, weighing dimensions of diagnostic value is not like weighing welfare against the health of the ecosystem, two goods which are very different in kind, or the welfare of one person against that of another, goods which, though similar in kind, are difficult to commensurate with much fineness of grain due to the lack of intersubjective comparability of different minds. All four dimensions of diagnostic value are similar in kind, conceptually connected with one another, and seem valuable for very similar reasons, so it is difficult to see why one of those dimensions should be overridingly more important than the others, or what conceptual obstacles there could be to making trade-offs between them.

Under those assumptions, the overall diagnostic value of options is measurable by a continuous real-valued function $f(\mathbf{p}, \mathbf{c}, \mathbf{a}, \mathbf{d}) : X \times Z \longrightarrow \mathbb{R}$, which is increasing in each dimension of diagnostic value, as well as separable across those dimension – i.e., $f(.)$ is a positively weighted sum of the four measures defined above. Since each dimension of diagnostic value has a natural 0 point, and since utility is cardinal, then if we further assume that diagnostic value along each of $\mathbf{p}, \mathbf{c}, \mathbf{a}$ is linear in the relevant cardinality-based measures of those dimensions of value, and make some conventional choice of unit for each of the four resulting dimension-specific scales of diagnostic value,[10] f is unique up to positive linear transformations. If diagnostic value along $\mathbf{d}$ is only measurable by a family of cardinal utility functions, rather than a unique cardinal utility function, then instead overall diagnostic value will be measurable by a family of f's, each unique up to positive linear transformation.

To move from f to an overall measure of the diagnostic value of an entire opportunity set, we might seek inspiration from the standard measures of opportunity set value. One naïve suggestion that quickly comes to mind is to rank opportunity sets by their most diagnostically valuable alternative, à la Pattanaik and Xu (2015). Thus we might say that A is more diagnostically valuable than B just in case the maximum over all $x \in A$ of $f(x, A)$ is at least as great as the maximum over all $x \in B$ of $f(x, B)$.

This suggestion is clearly incorrect, however. To the extent that it is attractive at all to rank opportunity sets by the value of their top-ranked option, this is because of the implicit belief that the instrumental value of an opportunity set is just the value of the outcome individuals achieve with those opportunities, and since a rational person will always choose the option they most prefer, it follows that the instrumental value of an opportunity set is the value of the outcome of the best option. Now, there is surely something right about the idea that the diagnostic value of an opportunity set has something to do with the diagnostic value of the outcome that individuals achieve with those opportunities, but it obviously does not follow from the assumption of rationality that individuals will always choose the most diagnostically valuable option: they may have other reasons for choice. A's most diagnostically valuable element may be more diagnostically valuable than B's most diagnostically valuable option, but if the remaining options in A are

[10] This assumption guarantees that the four scales are unique up to a common scalar multiplication, which matters, because otherwise f is only unique up to dimension-specific linear rescalings. One way of doing this is to assume, as seems plausible, that there is a least upper bound on how diagnostically valuable an option can be along any given dimension, e.g. there is a limit to how many new possibilities some choice can make you aware. Then we simply set, for each scale, the value of this least upper bound to 1, and this gives us a common unit of measure for all four scales.

not diagnostically valuable at all, but are attractive for other reasons, while the most diagnostically valuable option in B is also attractive for other reasons, the individual may extract much more diagnostic value from B than from A.

These last remarks suggest that an analogue to Vallentyne's proposal would make a much more appropriate measure of diagnostic value. Thus, if $q(x, A)$ is the probability that the individual will choose option x, given that A is their opportunity set, we can now propose that A is at least as diagnostically valuable as B just in case $\sum_{x \in A} q(x, A) f(x, A) \geq \sum_{y \in B} q(y, B) f(y, B)$ for every f in the family of admissible measures of overall option diagnostic value. I am interpreting q subjectively here, as representing the beliefs of some evaluator or modeller (or perhaps even the individual's own) about what the individual will choose. This interpretation has the benefit of being metaphysically innocent: as Inwagen (2000) and Buchak (2013a) have argued, to postulate objective unconditional probabilities of humans making particular choices seems to rule out the possibility of free will. Subjective probabilities, however, are un-mysterious: individuals are always trying to predict others' behaviour, and indeed sometimes their own.

Note that (to refer back to this section's opening remarks) an alternative way of constructing this expectational measure of diagnostic value would have been to first define dimension-specific value measures of the form $f_{\mathbf{i}} : X \times Z \longrightarrow \mathbb{R}$ and $F_{\mathbf{i}} : Z \longrightarrow \mathbb{R}$, where $\mathbf{i} = \mathbf{p}, \mathbf{c}, \mathbf{a}, \mathbf{d}$ and $F_{\mathbf{i}} = q(x, A) f_{\mathbf{i}}(x)$. We then say that $A \succsim B$ just in case the weighted sum of all the $F_{\mathbf{i}}(A)$ is at least as great as the weighted sum of all the $F_{\mathbf{i}}(B)$. This ranking rule is the same rule as the one I proposed.

An advantage of the expected value approach is that because each f is unique up to positive linear transformation, and therefore associates to each element in $X \times Z$ an absolute level of diagnostic value (starting from 0 diagnostic value), we have the flexibility to make our ranking of Z sensitive to the fact that different options within a single opportunity set may be unequally diagnostically valuable. Two opportunity sets, A and B, may have equally diagnostically valuable expected outcomes, yet if the variance in the diagnostic value of the options in A is much greater than the variance in the diagnostic value of the options in B, then the actual diagnostic value of the outcome the individual would achieve with A will tend to diverge much more from our expectations that the actual diagnostic value of the outcome the individual would achieve with B. Now, some level of risk aversion with respect to diagnostic value would justify ranking B over A in this case. And risk-aversion is well justified in some contexts: it makes the impact of policy choices more predictable, by minimizing the divergence between actual outcomes and expectations, and it minimizes the chance of poor diagnostic outcomes for individuals, which is a good thing if one thinks that duties of beneficence generally imply a greater interest in preventing bad outcomes than in promoting good ones.

Thus, one might propose that A is more diagnostically valuable than B just in case $\sum_{x\in A} q(x)\Phi(f(x,A)) \geq \sum_{y\in B} q(y)\Phi(f(y,B))$, where $\Phi(.)$ is an increasing, strictly concave transformation of f, say, the square root function. This transformed expected-value ranking is thus quite attractive: it induces a ranking over Z that is complete up to the level of completeness of the individual's preferences over preference learning experiences, it ties the diagnostic value of a set to the diagnostic value of the outcome we can expect individuals to achieve, and embeds a seemingly justified level of risk-aversion in its ordering.

Summary

I have tried in this chapter to draw attention to a neglected aspect of the value of freedom, namely the value opportunities have in generating knowledge which makes us better judges of value. Being a better evaluator is useful in helping us achieve better outcomes, but it is also good in itself. Diagnostic value is accordingly distinct in kind from ordinary prudential value, understood as the tendency of an option to satisfy one's desires, and from representative and symbolic value. Given that our preferences are often incomplete, that we may be uncertain about how they will change over time, and that we frequently lack information that is material to making good judgments, the value of having options that are diagnostically rich is not to be overlooked. Measures can be defined for each dimension of diagnostic value, three of which are guaranteed to be complete, and from these we can define plausible overall measures of both option and opportunity set diagnostic value.

****

This chapter ends our discussion of the opportunity aspect of freedom. In conclusion, I wish to offer some thoughts on how the intrinsic and instrumental value of freedom combine to determine the overall value of freedom. Presumably, the overall moral value of freedom will be positively related to both the intrinsic and instrumental value of freedom. Moreover, intrinsic and instrumental value are paradigmatically separable concerns: in every context where these distinctions matter, it is customary to separate one's assessment of how the intrinsic value of some activity (e.g. learning for its own sake) speaks in favour of it from one's assessment of how the instrumental value of said activity (e.g. learning because it's necessary to get a job) speaks in favour of it. Supposing this general principle applies here, the simplest view one can propose on the overall value of freedom is that it is some sort of convex mixture of the intrinsic and instrumental value of freedom – and it will therefore NOT necessarily be increasing in the extent of one's freedom, which is an interesting conclusion.

We are now ready to turn to the final theme of this thesis, concerning liberty, or the process aspect of freedom. So far, I've said quite a bit about the value of being free – of being *a free person* – so it is appropriate that we now turn to consider how it is that *respect* for free persons and their capacity for free choice constrains how we may interfere in their decisions about how to live their lives as they see fit.

Chapter 6

Liberty, Rights, and Social Choice

INdividuals have rights. These rights express our respect for them as free, autonomous beings, and constrain what inroads into personal liberty may be sanctioned by the pursuit of social welfare. That said, it speaks in favour of any course of action or public policy that it improves social welfare. Accordingly, I address two key questions in this chapter. How do rights and welfarist concerns interact to constrain social choice? And when can we say that one social state respects individual rights to a greater extent than another? Or in other words: how much protection does the process aspect of freedom require? And how can we assess when it is better protected? To answer these questions, I defend a novel choice theoretic approach to the representation of rights which allows us to define rights-based social choice criteria, and study how these criteria interact with standard welfarist criteria. I then operationalize this approach to define philosophically meaningful measures of the degree to which social states respect individual rights. These measures could be usefully applied to assess government or social policy performance on rights issues.

The point of departure for my argument is Sen's (1970) celebrated "Impossibility of a Paretian Liberal" result, in which Sen proposed a way of formalizing rights within a social choice theoretic framework, and showed that even the most minimal rights-based constraint was inconsistent with the weak Pareto principle, arguably the least controversial welfarist constraint. Sen's impossibility result vividly illustrates the conflicting demands which respect for rights and concern for welfare place on morality. However, his formalization of individual rights has been criticized. Rights, on his view, amount to a privilege of dictatorship over pairs of entire social states. Nozick (1974) formulated an influential critique of this conception of rights, showing that it quickly led to paradoxes, and so proposed that rights had to be conceived as protections against interference in one's choices, rather than dictatorship privileges. This critique inspired numerous economists (e.g., Sugden, 1985; Gaertner, Pattanaik, and Suzumura, 1992) to formalize rights through game forms, instead of axiomatic constraints on a social choice function, and this approach has gained acceptance among social choice theorists.

Though the game form approach has important virtues, I argue that it also has one important shortcoming: by abandoning the attempt to formalize rights as a constraint on permissible social choice, it does not capture the conception of rights Nozick deployed in his critique of Sen, and which is widely accepted among philosophers. This conception, I will show, can be better expressed in a social choice-theoretic framework, albeit one that differs from Sen's. My main goal in what follows is to show just how fruitful this analysis is: using the language of social choice theory allows us to talk about how many general kinds of rights people have (e.g., freedom of speech), as opposed merely to how many individually enumerable acts they have a right to perform, how severely these kinds of rights may be infringed, and what makes one social state more illiberal than another – none of which is possible under the game form approach.

This chapter runs as follows. In section 2, I contrast Sen's conception of rights with the game form conception of rights, and argue that while the latter improves on Sen's, it still has limitations, and it continues to run afoul of Nozick's critique. In section 3, I present a novel approach to the representation of rights in the language of social choice. This framework allows us to recover an analogue to Sen's original impossibility result, but also to present possibility results which capture standard deontological views about how rights constrain the pursuit of social welfare. In section 4, I move on to discuss the possibility of measuring how illiberal a social state is, and I show that my framework allows us to express a variety of desirable axiomatic conditions on a measure of illiberalism, and I present representation results for two such measures. The significance of these results is then discussed. A practical upshot is that these measures will enable us to improve our evaluations of countries' human rights performance. A theoretical upshot is that, with slight tweaks, these measures will also allow us to represent how strong an individual's deontic reasons to avoid performing some act are, providing a sort of representation result for a class of deontological theories.

6.1 Two formal conceptions of rights

6.1.1 Rights as decisiveness

Sen (1970, pp. 152-153) frames rights as constraints on a social decision function, which is a mapping from any set of individual orderings $\succsim_i$ over the set of possible social states X to a unique social preference relation $\succsim$ over X, the range of which is limited to those preference relations which specify, for every subset of X, a (not necessarily unique) "best" alternative. He assumes that individuals have "dictatorship privileges" over social states which differ only with respect to how they are situated. For example, if x and y differ only with respect to whether your bedroom walls are painted white or pink, *you* get to decide between x and y, and your decision is not to be susceptible to social censure.

From here, Sen (1970, p. 154) defines Liberalism, minimally, as the condition that there should be at least two individuals in society such that each individual is "decisive" over at least one pair of social states: an individual is decisive over a pair (x, y) iff, if that individual prefers x to y, then x is socially preferred to y, and vice versa if that individual prefers y to x. Sen's remarkable result is that this condition is logically inconsistent with two further apparently reasonable restrictions on the social decision function: Universal Domain, which requires that the domain of the social decision function include any profile of individual orderings, and Weak Pareto, which requires that if every individual prefers x to y, then so does society.

To see that these conditions are inconsistent, Sen asks us to imagine two individuals, Prude and Lewd, and three possible social states – a ("no one reads Lady Chatterley's lover"), b ("Prude reads Lady Chatterley's Lover and Lewd doesn't), and c ("Lewd reads Lady Chatterley's Lover and Prude doesn't"). Universal domain allows us to suppose that $a \succ_{Prude} b \succ_{Prude} c$ and that $b \succ_{Lewd} c \succ_{Lewd} a$. If the social decision to be taken is between a and b, then Prude should surely be decisive, and similarly, if the social decision to be taken is between c and a, then surely Lewd should be decisive. The Pareto condition implies that $b \succ c$, but Sen's condition of minimal liberalism implies that $a \succ b$ and $c \succ a$, and thus a cycle emerges, leaving no best alternative exists.

Commentators on this result agree that it shows that there is a tension between a respect for rights, on the one hand, and welfarism, on the other. However, many of those same commentators see a problem with the conception of rights suggested by Sen's result. As Gaertner, Pattanaik, and Suzumura (1992, pp. 166-167) and Sugden (1985, p. 218) show, if we try to formalize all rights as defeasible decisiveness powers, certain intuitively plausible assignments of rights become logically impossible. Sugden asks us to imagine Liz and Ken, who both write diaries; let pn be the social state in which Liz writes a political diary and Ken a nonpolitical diary, nn the social state in which neither writes a political diary, np the state in which Liz writes a nonpolitical diary and Ken writes a political diary, and pp the state in which both write political diaries. Presumably, Liz should be decisive between pn and nn, since the only difference between these two states is the content of *her* diary. For the same reason, she should be decisive between np and pp, and Ken should be decisive over pn and pp, and over nn and np.

However, Liz and Ken cannot logically be decisive over these pairs. Universal Domain allows us to suppose that $pp \succ_{Ken} nn \succ_{Ken} np \succ_{Ken} pn$, thereby implying that neither np nor pn can be socially preferred, and that $np \succ_{Liz} pn \succ_{Liz} nn \succ_{Liz} pp$, thereby implying that neither nn nor pp can be socially preferred. But then no social decision function exists which satisfies Universal Domain and respects Liz and Ken's rights. This seems to imply that for social choice to be possible either society must restrict what preferences individuals may express, or that Liz and Ken don't have a right to write what they like in their own diary. Either is chilling.[1]

6.1.2 Rights as game forms

The key problem with Sen's social choice approach was correctly diagnosed by Nozick (1974, p. 166): rights are not entitlements to the satisfaction of one's preferences over social states, but rather protections against interference in one's *choices*. In Liz and Ken's case, due respect for liberty requires that they both be allowed to freely decide what to write in their diary, and then, *que sera, sera...* If the outcome is not to someone's liking, that is unfortunate, but it does not imply any intrusion upon their liberty. Nozick's insight poses a serious challenge for the attempt to represent rights in the language of social choice theory, because the mappings from individual preferences to social preferences which are the bread-and-butter of social choice theory are typically not sensitive to choice behaviour.

These reflections pushed authors like Gaertner, Pattanaik, and Suzumura (1992) to propose that rights be defined as *game forms.* The thought is that different ways of organizing society can be represented by game forms, which are differentiated by what strategy profiles they assign to individuals. Of those available game forms, one of these game forms is assumed to be the "correct" game form, in the sense that you have a right to ϕ if ϕ is (or makes up part of) an available strategy to you in

[1] Gibbard (1996, pp. 400-402) defended a formalization of rights as decisiveness privileges that is consistent and does not run into the problem cited here, but he does so by adding a forfeiture condition. Roughly, if your exercise of your right would block the emergence of a Pareto superior outcome, your right is forfeited and imposes no constraint on the social decision function. However adding this condition does not improve things, because if rights are forfeitable in Gibbard's sense, then then the cycle between Ken and Liz is broken by *forcing* Ken to give up his right of decisiveness over pp and pn, thus implementing pn. But now Ken can no longer decide what to write in his diary, and his choice is dictated by Liz's preferences. Moreover, notice that if Ken is free to choose what diary to write, but is ignorant of what Liz will write, his maximin *and* maximax strategy is to write a political diary; likewise Liz's maximin and maximax strategy is to write a nonpolitical diary. It then follows that if Liz and Ken both follow their maximin-*cum*-maximax strategies, the social state which would result is np, not pn. Gibbard's approach therefore implies that allowing Liz and Ken to follow their optimal strategies – which leads to Liz's top-ranked outcome – will violate Liz's rights. Satisfying Liz's rights forces the implementation of her second-ranked outcome. This is counterintuitive.

that game form. Your rights are violated just in case the strategies that are in fact available to you do not match the strategies that would be available to you in the correct game form. What game form counts as the right one? Sometimes this is a matter of stipulation: for example, Sugden just assumes that Liz and Ken both have a right to write what diary they like, and so the right game form here is a normal form game with two players, each of whom has two pure strategies: "write political diary," and "write non-political diary." In Pattanaik and Suzumura (1996) and Suzumura and Yoshihara's (2006) approach, the right game form is the game form that society would prefer. On their approach, rights constrain social choice by forcing the selection of social states to be done by a two-step procedure: first society decides what game form should be used, out of all the feasible game forms that could be used, then the social state that is selected is simply the outcome of the game.

Of course, we don't have to do things this way. Most deontologists would probably object to allowing the content of people's rights be determined by the outcome of some social choice procedure. So instead we could simply assume that there is a privileged game form given to us by the true theory of rights, and then take it as a criterion of evaluation on social choice that it implements the game form that deviates the least from this privileged form. The true objects of social choice will then be game forms, or probability distributions over game forms, which actually seems like a plausible characterization of the empirical process of social choice: governments and other collective decision-making bodies never determine a complete social state, they only ever determine strategy profiles, making certain courses of action possible and others impossible for different individuals, and then individuals make the choices they make given the options they have.

So the game form approach has much to recommend it. That said, it isn't quite as expressive as one might like. What, after all, is a *right* in this framework? A right can't be an entitlement to having a particular strategy made available to us, at least in the way rights are usually conceptualized, because a strategy may involve doing a huge sequence of actions, all of which are not protected by the same rights. Suppose an available strategy to you in a game is "criticize the government, then join a golf club." The first of those two acts is protected by a right to free speech, but the second by a right to free movement and free association. So even if some game form accurately represents the current state of society, one can't read off this representation what kinds of rights people have (e.g., freedom of speech, association, movement, etc.). *A fortiori*, if some strategy we're entitled to isn't available, we can't tell which of these rights are violated nor the extent to which they are violated.

This matters. Some rights are more important than others: freedom of speech usually ranks higher, in most philosophers' hierarchies of rights, than freedom of commerce or freedom of automotive circulation (we're much more willing to fine people for their driving than for their speech). Some rights may also be unequally important to different individuals, and therefore the same restrictions on their rights may be easier or harder for them to bear. So even if you could measure "distance" from some ideal game form, this distance measure wouldn't tell you everything you need to know to accurately assess how well people's rights are respected by different actions and social arrangements. Likewise, it goes directly to the severity of a wrongdoer's crime which rights they violated, and to what extent, and this will not be revealed simply by counting how many strategies the wrongdoer makes unavailable. So while there's nothing incorrect about the game form approach, it's just a bit too coarse-grained to represent everything that is of moral concern with respect to rights. I therefore wish to propose a different representation, which is in some respects an extension of the game form approach, but is more fine-grained and expressive and, I hope, philosophically illuminating.

6.2 A new framework of rights

To motivate my framework, I will adopt a simple heuristic device. Suppose we place ourselves in the shoes of an impartial but sympathetic and liberal observer, who is tasked with advising society on what social outcomes it should aim to bring about. What would such an observer need to know in order to form their recommendations? For a start, the impartial observer will need to know how well everyone is doing in the various possible outcomes. But moreover, the liberal observer will need to know whether bringing about various different outcomes would violate anyone's rights. Since the observer is liberal, they will be exceedingly reluctant to advise any course of action that would require the infringement of individual rights. To evaluate any given social state, therefore, the liberal observer will need to know many things: how well is everyone doing? What rights do they have? Under what conditions are these rights infringed, and in which social states are these conditions met?

These last three questions raise an even more basic question: what it is, even, to have a right? Here I will rely on the classic Hohfeldian analysis of rights, as presented by Thomson (1992). Under this analysis, we can distinguish two kinds of "basic" rights: claims, and liberties. Claims are a three-place relation between one person, the claim bearer, a second person, against whom the first person holds her claim, and a proposition, which the second person is under a duty towards the first person of not allowing to be false. For example, to riff on Thomson, I might have a claim to a banana because you promised to give me one, a claim which is infringed if you eat the banana you owe me. Liberties are a more complicated many-place relation between one person, everyone else, and a proposition. Specifically, I am at liberty to make some proposition true if and only if I have no duty to anyone to ensure this proposition is false, and everyone else is under a duty to me to not prevent me from making it true if I try. I am at liberty to walk upon the beach iff no one has a claim against me that I don't, and this liberty is infringed iff someone imposes such constraints upon me as prevent me from walking upon the beach if I so choose.

In ordinary moral language, we tend to bundle liberties together in broad categories of *negative rights*: the classic liberal rights of freedom of speech, association, movement, conscience, etc. are all bundles of thematically related singular liberties to make particular propositions true, e.g., the liberty to make it the case that I say p and the liberty to make it the case that I say q are part of the bundle of liberties that constitute my right to free speech. And likewise we sometimes bundle claims together in broad categories of *positive rights*: it is usually held that the right to life has a positive aspect which entails a duty of rescue on those who can save you from death without excessive risk or cost; such a right entails all sorts of claims against all sorts of people that they not allow certain propositions to obtain, e.g., they owe me a duty not to let me drown if they pass me by a pond where I've passed out, a duty to pull me out of the way of a passing train if they can do so safely, etc.

We can model these ideas quite precisely. Given a finite boolean algebra of propositions Ω, define for every individual $i \in \mathbb{N}^+$ a unique schedule of rights as a finite family $R_i = \{R_{i1}, R_{i2}, ...\}$ with each $R_{ij} \subseteq \Omega$. The elements in each R_{ij} are propositions. Intuitively, they are the propositions which i either has a liberty to make true (and could in fact make true if no one prevented them from doing so), or has a claim against others to being true. I don't assume that every individual necessarily has the same number of rights, because under most theories of rights it is possible for individuals to forfeit certain rights and even bundles of rights (whether voluntarily or by violating the rights of others), and to acquire new rights (e.g., couples who get married for example give each other certain rights and duties). Now, x being a possible world, define $A^x_{ij} \subseteq \Omega$, intuitively denoting some set of propositions which are true or which the individual can make true at x, and which includes all the propositions from R_{ij} which the individual can make true at world x, or which they have a claim to being true and are in fact true at x. A right R_{ij} is infringed at world x iff R_{ij} is not a subset of A^x_{ij}.

With this, our liberal observer knows what rights individuals have, and under what conditions they are infringed. Now let the set X of all social states be the set of all pairs (I_x, A_x), where $I_x \subset \mathbb{N}^+$ is the population in possible world x, and A_x is a profile of sets $\{A^x_{i1}, A^x_{i2}, ...\}$, one for each $i \in I_x$. Since a unique social state (I_x, A_x) is associated to every possible world x, it is convenient to denote social states simply by the possible world x in which they occur. There are two important points to note here: first, because x is a possible world, the propositions that all individuals can *in fact* make true at x must be logically consistent with each other. Second, notice that that a profile of sets $\{A^x_{i1}, A^x_{i2}, ...\}$ will imply logically a profile of strategies in a game form, because the union across all j of A^X_{ij} specifies every proposition the individual can make true at x – so implicitly, the objects of social choice in my framework imply game forms.[2] However, these objects are also more informative, as under this way of specifying social states, a liberal observer *can* simply read off the specification of a social state how many individuals there are, which (if any) of their rights have been infringed, and (as we will see later) to what extent. Indeed, by the definition of what it is for a right to be violated, our liberal observer can identify the set $\mathcal{L}$ of all *liberal* social states as the set of all x such that for every $i \in I_x$, $R_{ij} \subseteq A^x_{ij}$. Appendix A gives a formal example of a liberal and illiberal state.

Our liberal observer now knows everything they need to know about people's rights in order to evaluate social states. The only thing they don't know is how well people are doing across the various possible outcomes. Obviously, for some outcome to be better than another for some individual i, i has to be a member of the two populations in terms of which these outcomes are defined. But supposing they are, let us suppose for ease of exposition that one outcome is better for i than another just in case i prefers it for its own sake (if you're unhappy with this, substitute any old theory of well-being instead). So for every $i \in \mathbb{N}^+$ let $\succsim_i$ be a preorder over the set $\{x|i \in I_x\}$ of worlds in which i exists. The symmetric and asymmetric parts of $\succsim_i$ are denoted $\sim_i$ and $\succ_i$.

With this fairly sparse vocabulary, we can easily recover an analogue to Sen's "impossibility of a Paretian liberal" result. Because Sen's result invokes Pareto considerations, which can only be applied to cases where the population of interest is fixed, we will confine our attention here to the set X' of all social states defined for some particular population $\overline{I}$. Define a social choice function as a function C which selects from every nonempty $C' \subseteq X'$, some nonempty subset of C',

[2] Specifically, a strategy is a maximal set of propositions that the individual can conjunctively make true. So, for example, if $A^x_{11} = \{a, \neg a\}$ and $R^x_{12} = \{b, \neg b\}$ are the two sets of propositions individual 1 is at liberty to make true, then individual 1 has four strategies available: $a\&b, a\&\neg b, \neg a\&b$ and $\neg a\&\neg b$.

intuitively, the outcomes which the liberal observer regards as permissibly selected by society. Finally, define a social choice rule as a mapping from the set of possible profiles of preferences over X' to the set of possible social choice functions. We can show that if there are at least two feasible social states, $x, y \in X'$ such that $x \in \mathcal{L}, y \notin \mathcal{L}$, then there is no social choice rule which satisfies the following three conditions.

> **(L) Liberalism**: $\forall x, y \in X'$, if $x \in \mathcal{L}$ and $y \notin \mathcal{L}$, then $y \notin C(\{x, y\})$.
>
> **(U) Universal Domain**: every logically possible n-tuple of individual orderings is included in the social choice rule's domain.
>
> **(P) Weak Pareto**: $\forall x, y \in X'$, if $\forall i \in \overline{I}, x \succ_i y$, then $y \notin C(\{x, y\})$.

Proof. Trivial. U requires there to be a social choice function C selecting some nonempty subset of $\{x, y\}$ when y is unanimously preferred to x. Since y is unanimously preferred, P implies that $x \notin C(\{x, y\})$, but since $y \notin \mathcal{L}$, L implies that $y \notin C(\{x, y\})$. Thus $C(\{x, y\}) = \varnothing$. Contradiction. □

For an intuitive illustration of this problem, consider the case of a master and their slave, or of a nation of oppressors and oppressed people, where the oppressed have been so beaten, broken, and brainwashed that they have come to prefer their servitude to their liberation. Of course, the oppressors prefer their victims to be servile and under their control rather than for them to be free. We thus have a situation where everyone in society prefers the social state where the slaves are unable to make true countless propositions which they've a right to make true. Weak Pareto implies that the social state in which the slaves remain slaves ought to be chosen over the state in which they are liberated and go free. Liberalism obviously implies the contrary. Hence, a contradiction.

The conflict between Pareto and liberalism may arise in other ways as well, as is well illustrated by the prisoners' dilemma. In the dilemma, two players have a right to make it true that they cooperate, or to make it true that they don't cooperate. Both players would prefer to be forced to cooperate than to be allowed to choose whether to cooperate or defect. Thus Pareto implies that society ought to force each player to cooperate so as to bring about the Pareto-dominant outcome, but Liberalism prohibits this very same choice. My Liberalism condition is thus hostile to the coercive engineering of Pareto superior outcomes. This case contrasts interestingly with the slavery case where it's clear that P should be violated rather than L (see e.g., Temkin 2015, 486), whereas in the prisoners' dilemma it seems plausible that L should be relaxed, at least provided the social benefits of forced cooperation were great enough.

L is too strict a condition, in my view, but it perfectly captures Nozick's intuition about how rights constrain social choice. By L, any social state which violates an individual's rights is eliminated, and the set of social states over which collective choice is legitimate is restricted to $\mathcal{L}$. Note that $\mathcal{L}$ may include many social states, and therefore L may leave multiple social states uneliminated, so even if we accept Nozick's totally uncompromising prioritization of rights, there is still *some* room left for social choice. We can, for example, define:

> **(P*) L-Strict Pareto**: $\forall x, y \in \mathcal{L}$, if $x \succsim_i y$ for all i, then $x \in C(\{x, y\})$, and furthermore, if, for some $i, x \succ_i y$, then $y \notin C(\{x, y\})$.

This condition is consistent with L and U. Additional restrictions may be imposed on the social choice rule, since many different social choice rules exist which satisfy U, L, and P*. For example, if welfare levels are interpersonally comparable, we can apply a choice-based analogue of the leximin rule to $\mathcal{L}$, and likewise with the elements in $X \setminus \mathcal{L}$; this would specify a unique choice function over X, one reminiscent of Rawls's two principles of Justice, with Liberalism capturing the demands of the first principle of justice, and leximin covering the demands of the difference principle.

My framework thus allows us to capture how, according to two very influential theories of justice, rights constrain social choice. In fact, a minor modification to my framework would also allow us to capture how, according to absolutist theories of deontology like Kant's, rights constrain individual choice: suppose C selects those outcomes which it is permissible for you to choose, and suppose we stipulate that if some individual i's j^{th} right is infringed in some outcome but you yourself are not responsible for this, then $A^x_{ij} = R_{ij}$. Then $\mathcal{L}$ will denote the set of all outcomes you can bring about without infringing anyone's rights. L thus prohibits you from violating anyone's rights, no matter the consequence, which is just what absolutist deontology requires.

The simple framework I have proposed for thinking about rights is general and expressive. For one thing, it is consistent with any substantive theory of rights: you may assign any individual you like whatever schedule of positive and negative rights your favourite theory requires (and those are the only two kinds on any view), and moreover you may choose whether or not to make these rights co-possible. This latter point is significant: authors disagree on whether rights must all be co-possible (Nozick, 1974; Steiner, 1977; Scanlon, 2008 argue they must, while Thomson, 1992 argues they are not). In my framework it is not required that the elements belonging to a right be logically independent from each other. It is possible within my framework for one individual to have a claim to some proposition whose negation another person has a claim to (of course, if rights aren't co-possible, then there are no liberal states). But obviously you can interpret schedules of rights so that they are all logically independent from each other, in which case rights will all be co-possible.

In addition, my framework is more expressive than the game form approach, seeing as allows us to read off the specification of the social state much valuable information which could not be be read off game forms: notably, how many bundles of rights people have (e.g., right to free speech, right to freedom of movement, etc.), whether any of these have been violated in any given social state, and, as we will see in the next section, how extensively these rights have been violated and how seriously morally wrong it is to violate them.[3]

A significant limitation of my approach is that it is synchronic, and assumes schedules of rights as exogenously given. It is therefore not sensitive to how the content of individual rights can change over time – in Hohfeldian terms, my analysis omits *powers*, which are abilities to cause changes in the content of one's rights or the rights of others. Criminals, by having committed crimes, are usually held to alienate many of their rights. For example, the freedom of movement of criminals may permissibly be curtailed in spectacular fashion through just punitive measures. A different assignment of rights is needed to represent the criminal's situation before and after the crime: "take a trip to Manilla" would be a protected proposition before committing a crime, but not afterwards. Ideally we would like an extension of the present framework that showed how acts by individuals in one period endogenously determines the assignment of rights in the next.

This limitation aside, the framework I've proposed seems adequate for the purpose of expressing liberal commitments in the language of social choice. In fact, as I will now show, it allows us to define measures of illiberalism.

6.3 Foundations for a measure of illiberalism

An impartial liberal observer's concern for rights is not exhausted by the question of whether a social state is liberal or not. They will also care about the degree of illiberalism of different social states. If there was no way for whomever they were advising to avoid infringing anyone's rights, the liberal observer would not bottom out of advice: even if we bracket all welfare concerns, they would advise minimizing, in some fashion, the extent of one's disrespect for the rights of others.

[3] Pattanaik (2008) presents a non-game form approach to the representation of rights that takes the idea of rights as protections on choice seriously, but his framework cannot express any of this either, not can it represent positive rights, and it implicitly assumes rights are co-possible and that everyone has only a single right.

If we are to capture the judgements of such an observer, we stand in need of some *measure* of illiberalism, one which will sort out which social states are more unchoiceworthy than which, from the sole point of view of concerns for individual rights.

What would our liberal observer's judgments of the degree of illiberality of any state be sensitive to? Obviously they will care about how many people's rights have been violated, and how many of each of these individuals' rights have been violated. They may also care about how rights violations are distributed across the population. This is all stuff they can read off the specification of the social state. But moreover, they will care about how *deeply* these rights have been violated, and how *valuable* these rights were. Intuitively, the fewer the propositions I have a right to make true that I actually can make true at will, the more my right has been violated, and you do me greater wrong by violating my right more deeply. Stealing my book for a year is a worse offence than stealing it for an hour. Likewise, some rights matter more than others: many authors have argued that negative rights prime over positive rights, and many recognize hierarchies among negative rights. Likewise, political and civil freedoms are typically held by many to be more important than economic freedoms.

6.3.1 Definitions

All of this can be precisely expressed in my framework. In appendix A I show how to construct a family of functions $F = \{f_i : X \longrightarrow \mathbb{R} | i \in \mathbb{N}\}$, unique up to a common affine transformation, with each f_i measuring the extent to which individual i's rights are disrespected. Each f_i is sensitive to the number of i's rights that are infringed, how deeply they are infringed, and how valuable the infringed rights are, and its uniqueness implies that we can make interpersonal comparisons in the degrees to which people's rights are disrespected. The process for constructing these functions is not complicated, but it is rather long and tedious, so I have decided to leave it only as an appendix. The interested reader may consult it to convince themselves that a philosophically well-grounded measure of the extent to which rights are disrespected can be constructed – I'm not simply assuming such a convenient structure exists! We can now associate to any outcome (I, A_x) an $|I|$-length row vector $\mathbf{x} = [\alpha f_i(x) \; ... \; \alpha f_{i'}(x)]$ with $\alpha > 0$ that summarizes how severely disrespected the rights of every individual in I are in outcome x.

6.3.2 Main results

I am in a position to present my main results. Let $g : \mathbf{X} \longrightarrow \mathbb{R}$ denote a measure of illiberalism, and consider the following conditions.

> **(A) Anonymity**: for any two row vectors $\mathbf{x}, \mathbf{y}$ of the same length, if there is a permutation $\sigma : \mathbf{x} \longrightarrow \mathbf{y}$ then $g(x) = g(y)$.

In other words, if the only difference between x and y is the identity of the individuals whose rights have been infringed, then we must regard x and y as equally illiberal.

> **(RI) Replication Invariance**: if $\mathbf{y}$ can be obtained by multiplying $\mathbf{x}$ by a duplication matrix, then $g(y) = g(x)$.

This condition requires that if we can get from x to y just by duplicating the population in x some whole number of times, duplicating every single rights violation in the process, then x is equally illiberal as y. Or in other words, illiberalism is relative to population size. An example will help illustrate. Consider the row vectors:

$$\mathbf{x} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}.$$

You can see that $\mathbf{y}$ is an replication of $\mathbf{x}$, in the intuitive sense that we've just copied and pasted twice, in turn, each column. And sure enough, you can get to $\mathbf{y}$ from $\mathbf{x}$ via a duplication matrix:

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix}.$$

> **(PD) Pigou-Dalton**: for any two x, y, if there is a pair i, i' such that
>
> 1. $f_i(x) + f_i(y) = f_{i'}(x) + f_{i'}(y)$
> 2. $f_i(x) < f_i(y) < f_{i'}(x)$ and $f_i(x) < f_{i'}(y) < f_{i'}(x)$
> 3. $f_{i"}(x) = f_{i"}(y)$ for all other $i" \in I$
>
> then $g(x) > g(y)$.

PD says that if x can be obtained from y by "transferring" rights violations from one individual to another whose rights (prior to the transfer) were overall disrespected to exactly the same degree, leaving everyone else's situation unchanged. The "transfer" makes the distribution of rights violations over these two individuals strictly more unequal than before. PD, in other words, encodes in $g(x)$ an aversion to strict increases in the degree of inequality with which rights violations are distributed across the population.

> **(PI) Pareto Illiberalism**: for any x, y, if $f_i(x) \geq f_i(y)$ for all i then $g(x) \geq g(y)$. Further, if there is an $i \in I$ such that $f_i(x) > f_i(y)$ then $g(x) > g(y)$.

Illiberalism increases in the degree to which each individual's rights have been disrespected, i.e., the more people's rights are disrespected, the worse things are

> **(C) Continuity**: $g^{-1}(V)$ is an open set in X for every open set V in $\mathbb{R}$, where $g^{-1}(V) \equiv \{x \in X | g(x) \in V\}$.

It's hard to read off C what, intuitively, it requires. But roughly, it says that there are no qualitative "jumps" in the value assessments encoded in $g(x)$, e.g., infinitesimal changes in how deeply a right is infringed can only make infinitesimal changes in the value of $g(x)$. Or in other words, evaluative judgements should not be hypersensitive to arbitrarily small differences in the non-evaluative facts they supervene on. In this case, the non-evaluative facts are the degrees to which each individual's rights are disrespected, and the identity and number of people in the population of interest.

> **(S) Separability**: for all x, y, z, zz with the same population size such that $f_i(x) = f_i(y) \iff f_i(z) = f_i(zz)$, and for any i such that $f_i(x) \neq f_i(y)$ it is the case that $f_i(x) = f_i(z) \iff f_i(y) = f_i(zz)$, it follow that $g(x) \geq g(y) \iff g(z) \geq g(zz)$.

This axiom requires that the assessment of two outcomes does not depend on individuals unaffected by the choice between these outcomes. An example again will help illustrate. Consider the two pairs of outcomes below:

$$\mathbf{x} = \begin{bmatrix} 0.5 & 0.4 & 0 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 0.45 & 0.4 & 0.1 \end{bmatrix}$$

$$\mathbf{z} = \begin{bmatrix} 0.5 & 0.8 & 0 \end{bmatrix}, \mathbf{zz} = \begin{bmatrix} 0.45 & 0.8 & 0.1 \end{bmatrix}$$

Separability implies that $g(x) \geq g(y) \iff g(z) \geq g(zz)$, because the only difference between x, y and z, zz is the degree to which the rights of the individual in the second column were infringed, but this individual is unaffected by the choice between x and y and by the choice between zz and z.

> **(SI) Scale Invariance**: if $\mathbf{x} = \alpha\mathbf{y}$ then $g(x) = g(y)$.

This final axiom says that if we rescale the functions which represent the overall degree to which each individual's rights are disrespected by a common factor, we don't change how illiberal the outcome is. As is made clear in appendix A, constructing the f_i requires a conventional choice of unit. Rescaling here amounts to changing the unit of measurement.

To see the motivation behind these conditions, let us place ourselves once again in the shoes of an impartial but liberal observer who is trying to assess how well individuals' rights are respected in states of varying sizes, and imagine $g(x)$ as a representation of our assessments. Seen in this way, conditions A and PI are *sine qua non* properties of $g(x)$. An impartial liberal observer would not care about the identity, as such, of those whose rights are violated, and they would judge things to be worse the more people's rights are disrespected.

SI and C are technical requirements on $g(x)$, rather than substantive ones, but they are innocuous. Regarding SI, it's clear that nothing essential hangs on our choice of unit for f_i. All that matters is that there is a common normalization across all i, because this captures the notion that everyone's rights matter equally, which is a foundational commitment of any liberal theory of rights, and that there is an upper bound on how deeply any right can be infringed, which is true by construction. This common normalization is preserved by common scalar multiplication. C, meanwhile, is attractive for guaranteeing that optimization problems have solutions. If the imperative to avoid infringing rights competes with the aim of promoting welfare, C guarantees that there is at the very least a boundary of efficient solutions, such that one cannot further either end without hampering the other. This would be a weak argument for it if there was any reason to think C imposed potentially distorting constraints on our impartial observer's judgements, but thankfully this isn't so. Our ordinary judgements about how well-respected individuals' rights are in different outcomes don't appear to be hypersensitive to arbitrarily tiny differences between those outcomes. Indeed, note that f_i is already a continuous function; it would be very surprising if our judgements about how severely an individual's rights may be disrespected were continuous, but not our judgements about how severely everyone's rights may be disrespected overall. C is therefore innocuous.

RI and PD are more substantive, but likewise very compelling. To appreciate RI, imagine that differently sized outcomes correspond to outcomes in differently populated countries. Our impartial observer shouldn't penalize some countries for having a larger population than others, because assessments of the degree of illiberality of a state should track something like how well that country is doing to protect the rights of its citizens, and such assessments must necessarily be relativized to population size. If there are ten times as many rights infringed in the US as in Canada, but there are also ten times as many people in the US, then, all else being equal, neither country has done better than the other in protecting its citizens' rights. And to see that an impartial liberal observer would respect the PD, it must simply be remembered that liberalism is inherently hostile to tyranny, in general, and to the tyranny of the majority, in particular. Majorities must not permitted to statute with impunity on the rights of minorities, so so we ought to penalize social states that marginalize some people for the benefit of others.

This leaves S, which is not, in my view, a necessary condition on $g(x)$, because whether or not it is desirable to impose it will depend on how exactly we understand the motivation driving our impartial observer's aversion to inequality. Our impartial observer will satisfy S if they don't care, as such, for how individuals fare with respect to each other in terms of the degree to which their rights are disrespected, but only about how each individual is doing, and they reach an overall assessment of the outcomes by somehow aggregating these individual judgements. How is such indifference consistent with a positive concern for the distribution of rights violations across the population? Well, the impartial observer's dislike for inegalitarian distributions might stem from their special concern for those whose rights are most disrespected in *absolute* terms. That is, if they believe that there is added value to alleviating by an increment of e the degree to which i's rights are disrespected when i's rights are more disrespected in absolute terms than when they are less disrespected in absolute terms (i.e., $f_i(x)$ is closer to 0), then they will favour more egalitarian distributions. This is arguably a reasonable attitude for a liberal: the reason I'm so concerned with human rights abuses in Myanmar has nothing to do with the fact that the rights of the Rohingya are more disrespected than those of the military head brass, but rather with how incredibly profoundly, in absolute terms, their rights are disrespected. There is also a good practical reason to impose S, which is that if you know that some number of people are sure to be unaffected by your decisions, you won't need to gather data on how disrespected their rights are, and you can confine your attention to only the slice of the population that will be affected. In practice, this will facilitate the computation of $g(x)$.

That said, in justifying PD I appealed to a sense that liberalism is hostile to the tyranny of majorities over minorities. And you might think that this hostility is fundamental, as opposed to being derivative of a special concern for the rights of the most disrespected. That is, you can reasonably argue that the best explanation for PD is that liberalism is intrinsically opposed to *social domination*: we don't want tyrants and oligarchs lording over subordinate populations of drones and servants. Domination is an inherently relational concept though – there is no domination without a dominator and a dominatee. So you simply can't come to a correct judgement about how illiberal a social state is without looking at how individuals do with respect to each other. To give an example, you might argue that what's so intolerable about slavery isn't merely how severely the rights of the slaves are infringed, it's that their rights are infringed as deeply as they are while the masters and other free persons enjoy the full gamut of liberal rights – in one respect it would be better if *everyone*'s rights were infringed as deeply as the slaves, for at least then there would be no such stark tyrannies. If you buy this line of argument, then our impartial observer will want to infringe S.

I don't wish to take a stand on whether to impose S or not. The philosophical argument for violating it is stronger, to my mind, but there is a reasonable philosophical argument for imposing it as well, and tractability considerations militate in favour of imposing it. Accordingly, I propose two measures of illiberalism that disagree only about whether to impose S. Thus I show:

Theorem 6.3.1. *$g(x)$ satisfies A, SI, RI, PD, MI, C, and S iff*

$$g(x) \geq g(y) \iff \frac{\sum_{i \in I_x} \gamma(f_i(x))}{|I_x|} \geq \frac{\sum_{i \in I_y} \gamma(f_i(y))}{|I_y|},$$

γ being any strictly increasing and strictly convex transformation function.

Theorem 6.3.2. *$g(x)$ satisfies A, SI, RI, PD, MI, C, and violates S iff*

$$g(x) \geq g(y) \iff \frac{\sum_{i \in I_x} \beta_i \cdot f_i(x)}{|I_x|} \geq \frac{\sum_{i \in I_y} \alpha_i \cdot f_i(y)}{|I_y|},$$

where $\{\beta_i\} = \{\alpha_i\} \iff |I_x| = |I_y|$, all the β_i and α_i are positive and can be arranged in a strictly increasing order, such that a person whose rights are strictly more disrespected gets a greater factor β_i or α_i.

These results follow from very similar results proved elsewhere by Atkinson (1970), McCarthy (2018), Adler (2018), and Fleurbaey (2018). Sketches of the proofs are in the appendix.

We now have two well-behaved measures of illiberalism that capture a wide range of liberal commitments, but we're not done yet. I opened this chapter with the promise of offering a way of representing the strength of an *individual agent's* reasons not to perform some act, out of respect for the rights of others. Yet the heuristic I've used to justify the conditions above appeals to an impartial observer. Under a deontological theory of morality, there are no doubt contexts in which it is appropriate for an individual agent to take an impartial point of view on the world and survey, as it were, how well-respected individuals' rights are in different societies (e.g., this might be important if you're trying to decide in which society you'd like to live, or which society you would like to see brought about). But the impartial observer is really a consequentialist sort of daemon, and we cannot substitute their judgements for our own in most ordinary decision-making contexts.

There are two key reasons for this. First, deontology takes very seriously the distinction between the rights that *I* infringe and the rights that *others* infringe – a distinction that is not salient to the impartial observer, since they are precisely just an observer, not an agent – and they want to make room for permissible partiality. That is, in general, we have much greater reason to *avoid* infringing some rights than we have to *prevent others* from infringing those same rights. To be sure, if there were enough sufficiently weighty rights we could prevent others from infringing sufficiently deeply, then our reasons to prevent the rights violations might swamp our reasons to avoid comparatively much more minor infringements, but it takes some doing. Likewise, a mother is permitted to care more about protecting the rights of her child than about protecting the rights of her sister-in-law's third cousin once removed. This is all by way of saying that if you were to interpret $g(.)$ as representing the weight of *your* reasons not to bring about some outcome, purely out of respect for the rights of others, anonymity is a singularly implausible condition.

Second, while it makes sense to relativize illiberality to population size, it does not make sense to relativize the absolute weight of your deontic reasons to population size. This can be brought out with a pair of examples. Suppose you were choosing which country to move to: you would surely care about the *rate* of disrespect for rights in that country more than the *absolute extent* to which rights are disrespected. The United States are more populous than North Korea, so it's not impossible to imagine they infringe the rights of more of their citizen than North Korea does, just because there are so many more Americans than North Koreans, but relative to population size there is clearly more respect for rights in America. Wouldn't you rather live in the US? But suppose you face a choice between violating no one's rights and violating some number of people's rights; would the weight of your reasons not to infringe the many's rights diminish – i.e., would it make it easier to justify – if the total population were larger? Obviously not.

These remarks point to a simple way of modifying the measures derived earlier so that they measure instead the weight of our deontic reasons not to perform acts which are the bringing-about of various outcomes. Instead of averaging sums of weighted $f_i(.)$, we take the sum *tout court* of weighted $f_i(.)$, where the weights reflect not simply inequality aversion, but also the extra weight that the individual is permitted to place on the rights of those close to them, and the extra weight they must place on those rights which they would be personally responsible for infringing if they chose some particular course of action. Thus, under an "egalitarian" construal of what's wrong with inequality, the weight of an individual k's deontic reasons not to cause x can be represented by:

$$g_k(x) = \sum_{i=1}^{n} \beta_{xki} \cdot f_i(x),$$

where the weighing parameter is indexed to individual k, to allow for differential weighting based permissible partiality, and to the outcome x, to allow for differential weighting based on whether k is personally responsible for infringing individual i's rights by causing x, and to individual i, to ensure that, all else being equal, individuals whose rights are more disrespected get weightier β-parameters.

6.4 Summary

What is the significance of all these results? In an interview by J. S. Kelly, Kenneth Arrow confessed himself unhappy with the discourse of rights: it's not that the vocabulary of rights isn't often very appealing, but he worried that you can always make rights say what you want them to, and so he concluded that until "somebody produces a logic of rights in terms of which we can *argue*, I really find the whole issue is unfocused" (Sen, Arrow, and Suzumura, 2011). This chapter is my answer to Arrow's demand.

I've put forward a new way of formalizing rights. As we've seen, my approach is more general than other approaches, being capable of representing both positive and negative rights, and being consistent with any substantive theory of rights. In particular, it allows us to define conditions on permissible choice that perfectly capture Nozick's intuition about how rights constrain choice. But the main takeaway point is just that, within the basic framework I presented in hand, we can express in social choice-theoretic language a wide range of liberal – and, more generally, deontological – commitments, including commitments against social tyrannies of the majority, commitments to the equal value of everyone's natural rights and of everyone's interests in having their rights respected, and more. My results give deontology a proper "seat at the table" of social choice theory, when until now deontology and social choice theory have been to each other as two ships passing by in the night. All taken together, this is proof of concept, I believe, for my proposed way of fitting rights inside social choice theory.

Moreover, we have seen that the interaction of these commitments, taken together with a technical assumption (C), has a very specific structure which can be conveniently summarized in functional forms that are familiar in welfare economics and normative ethics. This has an immediate theoretical payoff. Section 3 showed how, in my framework, rights constrain social and individual choice for absolutist theories of deontology: the non-violation of rights always dominates. But with the two measures of illiberalism proposed in section 4 in hand, we can study in the language of social choice theory how the package of moral commitments that is summarized by the $g(.)$ or $g_k(.)$ of our choice interacts with other sorts of moral commitments, such as welfare-consequentialist commitments, which can themselves be represented as functions from X to the reals. This opens up an interesting research program for moderate deontologists who believe that our deontic reasons to respect rights can be traded off against our axiological reasons to promote welfare. Since both kinds of reasons can both be represented by continuous real-valued functions over a common set of outcomes, we can combine them in various ways (e.g., linear combination) and study how thus combined they constrain permissible choice.

My results also have significant practical payoffs: many think tanks and NGOs have an interest in measuring how well individual freedom is respected in different countries. The Cato and The Fraser Institute (2019) have for many years produced a Human Freedom Index, self-consciously grounded in a conception of freedom owing to Locke (1948) and Berlin (1969), which ranks countries by the extent of the economic and personal freedoms they offer, based on 76 sub-indices which we then average over. Each indicator is ranked on an intuitive scale from 0 to 10, with 0 indicating no freedom along that indicator and 10 indicating maximal freedom, then they average out the scores, weighing the overall economic freedom score and the overall political freedom score equally. Other measures of human rights have been proposed in the political science literature, all using a similar methodology. The most influential measure of human rights in political science is the "Gastil scale," which involves simply drawing up a checklist of political and civil liberties and scoring countries from 1 to 7 on each list depending on how poorly, in the evaluator's judgement, they fill out the checklist (see e.g., Gastil, 1990; Muller and Seligson, 1987; Foweraker and Landman, 2000; Landman, 2004; Stepan and Skach, 1993).

These scoring rules can be quite illuminating, but ultimately the score just reflects an evaluator's ordinal intuitive judgements of how well countries do with respect to different indicators. So, fundamentally, it isn't clear how the principles which motivate the scoring rules actually constrain their computation. Moreover, averaging or summing over intuitive scores, as these indices require, is also dubiously meaningful, because these numbers merely represent an ordinal ranking of countries, and because the different indicators measure quite different things – e.g., it's not clear what it means to average over the integrity of the legal system and the absence of restrictions on freedom of movement. One worries it's a bit like averaging my weight in Kg with my body temperature in degrees Celcius: both are important health variables, but they cannot be meaningfully averaged. In contrast, with my measures of illiberalism, liberal commitments axiomatically constrain the computation of the indices, and guarantee the meaningfulness of comparisons in overall degrees of illiberalism. My framework therefore opens up the possibility of constructing better-founded indices of liberty.

Chapter 7

Uncertain Liberty

I Promised in the previous chapter to discuss how much protection the process aspect of freedom requires, i.e., how strong are the prohibitions rights impose on interference in the personal liberty of others. However, the possibility results I presented paint an uncompromising picture of liberty's priority over welfare. Liberalism and the limited Pareto principle prohibit any inroads into the protected sphere of individual liberty for the sake of improving social welfare. *Fiat Justitia, pereat Mundus.* This picture accords with general principles of constitutional design. We generally *want* constitutions to place hard limits on government power by enshrining certain rights and freedoms that no lawful act may breach. And it finds support from figures as divergent on most other matters as R. Dworkin (1977) and Nozick (1974), who defend the view of rights as, respectively, trumps against competing interests, and side-constraints on permissible choice. This is not surprising, as the framework I developed was designed to capture this Nozickean intuition.

Clearly, this picture clashes with common sense. It strains our sense of empathy and moderation to accept that no inroads into protected spheres of individual liberty, however tiny, may be sanctioned to avert any amount of human suffering, however acute and widespread. Nozick himself expressed discomfort at this thought, and many deontologists respond to these sorts of cases by introducing threshold views of the constraining power of rights, according to which rights can permissibly be infringed if the consequences of respecting them are sufficiently horrible (see, e.g., Thomson, 1984; Moore, 1989; Nagel, 1972). I am sympathetic to such views, and so, for the final chapter of this thesis, I propose to defend a moderated vision of how rights constrain social and individual choice.

Not everyone recognizes the need for moderation in the defence of rights. There are absolutists who insist that the world must sooner perish than we violate deontological constraints (Ellis, 1992; Alexander, 2000). I do not wish to beg the question against the absolutists. And so, rather than start with a positive argument for moderation, I propose to first show that the strict Nozickean vision enshrined in

Liberalism, which is defined in a context where the consequences of our choices are known to us, cannot be coherently extended to the context of uncertainty without undermining absolutist starting points. From this initial argument I build a case for moderation on premises that should be acceptable even to absolutists, and I show how, with the measures of illiberalism developed in chapter 6 in hand, we can weigh the right against the good in a way that is attractive.

My argument is structured as follows. In section 1 I extend the basic vocabulary developed in chapter 6 to the context of uncertainty, and I argue that the most natural extension of Liberalism to this context takes the form of an Extended Liberalism (EL) axiom which encodes a strict preference for any act guaranteeing a liberal social state to any act which does not. I then lay out a few impossibility results, showing that any social choice rule satisfying Extended Liberalism and two further conditions must violate non-atomicity and the sure-thing principle. Violating these axioms comes with well-known costs. This is an initial blow against absolutist deontology, but not a decisive one, since absolutists may be willing to pay such pragmatic costs if it purchases them a strong set of rights.

However, I argue in section 2 that EL contradicts what absolutists like Nozick themselves say about about the permissible exercise of our right to punish, given that we must rely on methods for identifying wrong-doers that necessarily carry the risk of false convictions. Indeed, under uncertainty, EL implies that such rights are *de facto* never permissibly exercised – a much more bitter pill to swallow for absolutists, since it largely strips our rights of value, and makes it harder to justify violating Savage's axioms. Making room for the permissible exercise of these rights requires relaxing EL, but then freedom loses its absolute priority over personal good in the transition from certainty to uncertainty. And if we are prepared to part with absolutism under uncertainty, then I suggest we face no pressure to accept it under certainty either. I thus move in section 3 to defend general principles for decision-making under certainty and uncertainty that encode a more moderate picture of the priority of personal liberty over social welfare. I then argue in section 4 that this new framework defuses standard objections to moderate deontology, notably the objection that it leads to strange verdicts in the neighborhood of the thresholds.

7.1 The framework of liberty extended

As in the previous chapter, we start with a finite boolean algebra of propositions Ω, and define for every individual $i \in \mathbb{N}^+$ a unique schedule of rights as a finite ordered set $R_i = \{R_{i1}, R_{i2}, ...\}$ with each $R_{ij} \subseteq \Omega$. The elements in each R_{ij} are propositions, intuitively, the propositions which i either has a liberty to make true, or has a claim against others to being true. Now define $A^x_{ij} \subseteq \Omega$, which includes all propositions from R_{ij} which the individual can in fact make true in outcome x, or which they have a claim to being true and are in fact true. A right R_{ij} is infringed iff R_{ij} is not a subset of A^x_{ij}.

The set X of all outcomes is the set of all pairs (I_x, A_x), where $I_x \subset \mathbb{N}^+$ is the population in outcome x, and A_x is a profile of sets $\{A^x_{i1}, A^x_{i2}, ...\}$, one for each $i \in R_I$. The set L of all *liberal* social states is the set of all (I_x, A_x) such that for every $R_i \in R_I$, $R_{ij} \subseteq A^x_{ij}$. As before, since exactly one pair (I_x, A_x) is associated with every possible world, it is convenient to denote outcomes simply by the possible world in which they occur. In liberal social states, no one's rights are infringed. Of course, if someone's rights are infringed in some outcome x, whether you would be the one to infringe those rights by bringing x about, or whether you would simply be failing to prevent others from infringing them, makes a difference for whether it would be permissible for you to bring about x. So let us simply stipulate that if some individual i's j^{th} right is infringed in some outcome but you yourself are not responsible for this, then, from your point of view, $R_{ij} \subseteq A^x_{ij}$. Under this interpretation, $\mathcal{L}$ is the set of outcomes you can bring about without infringing any rights.

There are different decision theories one can adopt to extend this basic framework to the context of uncertainty, but for ease of exposition I will adopt that of Savage (1954). On this framework, the primitives are states of the world, which are the features of the world over which we have no control, and consequences, which are the desirable or undersirable states of affairs that we can bring about through our actions. The consequences of our actions depend on the states of the world, and we are uncertain about which state of the world is the actual state of the world. Thus we begin by defining a partition S of states over the set of possible worlds, a set X of outcomes (defined as above), and a set Γ of *acts*, which are functions from S to X. Subsets of S are called events. We then define, for every individual i, a preference relation $\succsim_i$ defined over Γ. If this preference relation satisfies all of Savage's axioms, then their preferences over acts can be represented by a utility function unique up to positive linear transformations, and their beliefs about the state of the world can be represented by a unique probability distribution p_i defined over S.

The aim of the game now is to provide restrictions on a moral choice rule which maps profiles of individual preferences onto a choice function C which selects, from any set $A \subseteq \Gamma$, a nonempty subset of A, intuitively denoting those acts which are permissibly chosen from A. In chapter 6, I introduced the axiom of Liberalism, which encodes the Nozickean view of rights as side-constraints on permissible choice by prohibiting illiberal social states from being chosen when liberal states are available – that is, by prohibiting any rights infringements when it is instead possible not to infringe any rights. This intuition can be generalized to our present context by the following criterion:

> **(EL) Extended Liberalism**: for all $f, g \in A \subseteq \Gamma$, if $f(s) \in L$ for all non-null s, but there is a non-null s such that $g(s) \notin L$, then $g \notin C(A)$.

In other words, an act entailing the possible violation of someone's rights is not permissibly chosen over an act guaranteeing the non-violation of anyone's rights – morality prohibits gambling with people's rights when you don't have to. This condition entails the more limited condition of Liberalism, for in the degenerate case of certainty where $f(s_k) = x \in L$ and $g(s_k) = y \notin L$ for all $s_k \in S$, EL requires you to choose f and not g, which is just what Liberalism requires. EL receives direct support from Nozick's (1974, p. 106) own view on how epistemic considerations weigh on permissible action:

> If doing act A would infringe Q's rights unless condition C obtains, then someone who does not know that C obtains may not do A. Since we may assume that all know that inflicting a punishment upon someone infringes his rights unless he is guilty of an offense, we may make do with the weaker principle: if someone *knows* that doing act A would infringe Q's rights unless condition C obtains, then someone who does not know that C obtains may not do A. Weaker still, but sufficient for our purposes, is: If someone knows that doing act A would infringe Q's rights unless condition C obtains, he may not do A if he has not ascertained that C obtains through being in the best possible position for ascertaining this. (This weakening of the consequent also avoids various problems connected with epistemological skepticism.)

Under its two stronger formulations, Nozick's view straightforwardly entails EL. The weaker version also entails EL, subject to a small modelling choice. Nozick's third principle seems motivated by the thought if you've ruled out all possibility that ϕ-ing would infringe anyone's rights, *barring consideration* of unreasonable sceptical scenarios, then you've done your due diligence – no one could blame you for doing what was sure beyond a reasonable doubt to infringe no one's rights. So if in specifying S, we leave out any states describing absurd sceptical scenarios, Nozick's weak principle entails EL.

Extended Liberalism also flows from a standard Kantian picture of deontological duties. In the view Kant expresses in the Groundwork, duties and prohibitions apply primarily to maxims – or as Tennenbaum (2017, p. 691) puts it, "to what we can will directly (rather than by means of willing something else)." He goes on: "we can will directly only what we know that we can do by (in) willing it. The requirement to keep one's promises applies to an agent only insofar as she knows that by doing certain things, she'll be keeping her promise." Keeping this in mind, suppose that the antecedent condition of EL is satisfied. Then, under Tennenbaum's reading of Kant, the duty not to infringe anyone's rights applies to you, because you have an available alternative, f, which you *know* does not violate anyone's rights, thus you can will to violate no one's rights by willing f. Of course, you have another alternative, g, which you do *not* know will fail to violate anyone's rights. Is it permissible to choose g? Not as I understand Tennenbaum: since you do not know that by choosing g, you will violate no one's rights, you cannot will to violate no one's rights by (in) willing g. You would therefore fail to comply with Kantian obligation to *will* that you violate no one's rights by choosing g. And since I am assuming that the obligation to violate no one's rights is absolute, the fact that choosing g fails to meet this obligation entails that g cannot be chosen. The consequent of EL is therefore satisfied. Thus, Kantianism also leads us to EL. And EL is quite a restrictive condition, as it quickly generates impossibility results. Consider:

> **(U) Universal Domain**: every logically possible set of n individual preorders is included in the social choice rule's domain.
>
> **(LEAP) L-Ex Ante Pareto**: for any two $f, g \in A \subseteq \Gamma$ such that EL does not imply $f \notin C(A)$, if $f \succ_i g$ for all $i \in \mathbb{I}$, then $g \notin C(A)$
>
> **(P6) Non-atomicity**: If $g \notin C(\{f, g\})$, then for any $x \in X$, there is a finite partition $\{E_1, E_2, ..., E_m\}$ of S such that:
>
> - $f'(s_k) = x$ for any $s_k \in E_k$, but $f'(s_k) = f(s_k)$ for any $s_q \notin E_k$,
> - $g'(s_k) = x$ for any $s_k \in E_k$, but $g'(s_k) = g(s_k)$ for any $s_q \notin E_k$,
> - and $g \notin C(\{f', g\})$ and $g' \notin C(\{f, g'\})$.
>
> **(P2) Sure-Thing Principle**: if f, g, f', g' are acts such that:
>
> - $f(s) = g(s)$ and $f'(s) = g'(s)$ for all $s \in E \subset S$,
> - $f(s) = f'(s)$ and $g(s) = g'(s)$ for all $s \notin E$,
>
> then $f \in C(\{f, g\})$ if and only if $f' \in C(\{f', g'\})$

U requires that individuals be allowed to have what preferences they like. LEAP expresses the very weak unanimity condition that it is impermissible to implement a course of action which is universally dispreferred to an alternative that is itself not excluded by EL. P6 is one of Savage's axioms (only slightly reformulated here using a choice function rather than a preference relation), and intuitively it demands that there be no outcome *so* good or *so* bad that it swamps the improbability of any given event. That is, if f is already more choiceworthy than g, then f remains more choiceworthy even if we add x as an additional possible consequence of f, provided that the chance of x's occurrence is sufficiently small. The intuition behind Savage's famous sure-thing principle (likewise slightly reformulated) is that if two acts generate the same outcome in one event E, but a different outcome if E fails to materialize, then if one of those two acts is more choiceworthy than the other it is because it generates a better outcome in the event that E does not occur. Or to put things slightly differently, the desirability/choiceworthiness of some act is increasing in the desirability/choiceworthiness of each of its possible outcomes, if these could be brought about directly.

It is easily shown that no social choice rule exists which satisfies EL, U, LEAP, and P6. Let f, g be two acts such that $f(s) \in L$ for all $s \in S$ and $g(s) \in L$ for all $s \in S$. Universal Domain requires there to exist a social choice function for individual preferences for which $f \succ_i g$. LEAP then requires $g \notin C(\{g, f\})$. Now let x be an illiberal outcome. Non-atomicity requires that for any $x \notin L$ there is an event E and an act f' such that $f'(s) = x$ for all $s \in E$, $f'(s) = f(s)$ for all $s \notin E$, and $g \notin C(\{g, f'\})$. But since $g(s) \in L$ for all non-null $s \in S$ and $x \notin L$, Extended Liberalism implies $f' \notin C(\{g, f'\})$. Contradiction.

Likewise, no social choice rule exists which satisfies EL, U, LEAP, and P2. To see this, consider the following variant on the Allais paradox:

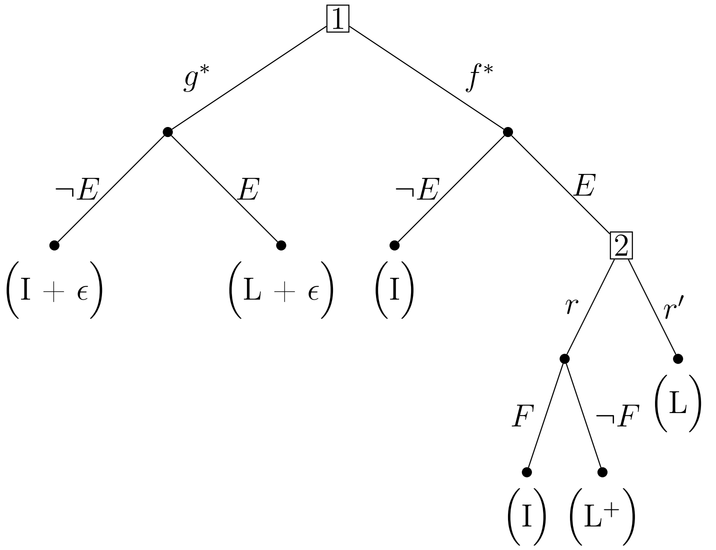
Table 1	E		$\neg E$
	F	$\neg F$	
f	Illiberal outcome	High-welfare liberal outcome	Illiberal outcome
g	Liberal outcome	Liberal outcome	Illiberal outcome
f’	Illiberal outcome	High-welfare liberal outcome	Liberal outcome
g’	Liberal outcome	Liberal outcome	Liberal outcome

Universal domain requires there to be a social choice rule for individual preferences for which $f \succ_i g$ for all $i \in \mathbb{I}$. Extended Limited Weak Pareto thus requires $g \notin C(\{g, f\})$. Since f agrees with g and f' agrees with g' in event E, and f agrees with f' and g agrees with g' in event $\neg E$, the sure-thing principle implies that $g' \notin C(\{f', g'\})$. But since g' generates a liberal outcome in every possible state of the world, Extended Liberalism requires that $f' \notin C(\{f', g'\})$. Contradiction.

Conditions U and LEAP are innocuous. A hard-line deontologist must therefore violate the Savage axioms if they wish to preserve EL. The cost of violating Non-atomicity is quite steep. It implies that morality requires us to pay *any* price for the certainty that we will infringe no one's rights. Thus even if you believe that the likelihood you will infringe someone's rights by ϕ-ing is astronomically small, you must be willing to pay *any* price to bring that chance down to 0. Indeed it implies we should be willing to pay as much for a great increase in our confidence that ϕ-ing will infringe no one's rights as for an infinitesimal increase in our confidence, provided the increase moves us to certainty. For those, like myself, who think that one good reason to reject consequentialism is that it is too demanding of agents, requiring excessive sacrifices in the pursuit of the impartial good (see also Williams, 1974; Lenman, 2000; McElwee, 2017), this is an embarrassing conclusion. And we're not out of the woods, because violations of the sure-thing principle don't come cheap either.

The standard pragmatic argument for obeying the sure-thing principle is that agents who violate it are diachronically inconsistent, and will sometimes choose courses of action which, by their own lights, are not as choiceworthy as their alternatives. In our context, the violation of the sure-thing Principle implies that we expose ourselves to being goodness-pumped: we will pay to do what we could have done without paying. To see this, consider the following choice tree, where square nodes represent choice nodes, where the agent decides whether to go left or right, and round nodes represent chance nodes, where nature decides whether we go left or right. Further, I let L denote a liberal outcome, L^+ denote a high-welfare liberal outcome, I denote an illiberal outcome, and ϵ some small benefit the agent could provide to themselves or to someone else. E and F denote probabilistically independent events, such that $p(E) = 0.11$ and $p(F) = 10/11$.

The decision situation in this tree is the following: at $t = 1$, the individual can choose between doing g^* or f^*. If the individual does f^* and E turns out to be the case, this puts them in a later position to choose between doing r or r', thus giving the agent three strategies to choose from: $g^*, f^* + r$, and $f^* + r'$.



Notice that none of our three strategies are ruled out by EL, since none is guaranteed to infringe no individual's rights – and in particular, at the first choice node, neither g^* nor f^* is ruled out by EL. Further, observe that acts $f^* + r$ and g^* have the same payoff as acts f and g from table 1, plus a small benefit of ϵ in the case of g^*. We know that $g \notin C(\{g, f\})$, but if ϵ is small enough, people's preferences will not reverse. A reapplication of LEAP therefore yields $g^* \notin C(\{g^*, f^* + r, f^* + r'\})$. But supposing the agent chose f^* at the first choice node and E turns out to be the case, then they arrive at the second choice node, where it is not permissible for them to choose r instead of r', because, unlike r, r' is an act guaranteed to infringe no one's rights. Since it is only rational to choose a strategy which it would later be rational to pursue if one had the choice not to, it follows that $f^* + r \notin C(\{g^*, f^* + r, f^* + r'\})$. Thus the only strategy that is left uneliminated is $f^* + r'$.

But this is an irrational strategy. Whether or not E obtains, g^* is more choiceworthy than $f^* + r'$: $f^* + r$ will turn out to infringe no one's rights if and only if g^* would also turn out to infringe no one's rights, but at least g^* is guaranteed to be of extra benefit to someone. Our reasons for choice become internally inconsistent: we are adopting a strategy we *know* we will be unable to commit to on the grounds that, assuming we *will* later commit to it, this strategy is more choiceworthy than its alternative.

There are two standard solution concepts for dealing with this sort of diachronic inconsistency. The first is Resolute Choice, which instructs agents to stick to the strategy they adopted at earlier choice nodes, even if by the time they reach later choice nodes, they no longer judge this strategy to be optimal. In our context, this would mean choosing to gamble with the rights of others by choosing r instead of complying with one's duty to infringe no one's rights. This is not an attractive solution for someone with absolutist sympathies, which leaves us with the second solution concept, Sophisticated Choice. Sophisticated choosers appreciate that not all strategies that are open to them at the first choice node in a decision tree are ones that they can follow through on, so they limit their attention to those strategies which it will be permissible for them to follow through on at every subsequent choice node. More precisely, assume the longest branches in some tree have k choice nodes. A sophisticated chooser decides what the best choice is at node k, and assumes no other choice is possible once they reach k, then works backwards until they reach the root choice node. In our context, since $f^* + r$ is not a strategy which it is permissible to follow through on at the second choice node, sophisticated choice implies that only g^* and $f^* + r'$ are permissible strategies, and between those two strategies, LEAP requires choosing g^*.

Sophisticated choice immunizes an agent obeying EL and LEAP against certain displays of diachronic inconsistency. Unfortunately, there are well-known results showing that sophisticated choosers will choose dominated strategies in other sorts of cases (Rabinowicz, 1995; Briggs, 2015), and that moreover they will sometimes pay to avoid free information that is relevant to their decision about what to do (Al-Najjar and Weinstein, 2009).[1] So the threat of irrationality looms large, and this gives us initial reason to be skeptical of absolutist deontology. Morality is very dearly bought if it comes at the expense of rationality. That said, the pragmatic costs of irrationality are paid in the coin of welfare, not the coin of failing to comply with duty, so the price of imposing EL may well be worth the prize for a really hard-nosed deontologist, as it extends absolutist commitments to the context of uncertainty. However, as the next section shows, absolutism has additional costs, and these are paid in the currency of deontology.

7.2 A dilemma for absolutism

In addition to having rights against interference, individuals are usually held to have the right to enforce those rights against others. Enforcement rights include the right to defend one's rights against intrusion by others, as in cases of legitimate self-defense, and to seek redress for past wrongs, for example by seeking to punish wrongdoers or forcing them to issue apologies, compensate their victims, or serve their community as repayment.

[1] Rabinowicz (1995) defends a third solution concept to the problem of dynamic inconsistency, wise choice, which essentially combines features of sophisticated and resolute choice. Wise choice avoids dynamic inconsistency by sometimes choosing resolutely (Peterson and Vallentyne, 2018, p. 68), which remains as unattractive in our context as before. Likewise, Buchak (2013b) has argued that the negative pricing of information can be rationalized as a strong desire to avoid being misled, but this is hotly contested.

According to the natural rights tradition that follows Locke (1948, 2nd Tr. §7), all individual have the right to defend themselves and others from aggression, and to punish rights-violators and force them to compensate their victims, if compensation is possible. We lose these rights only if we voluntarily alienate them by infringing the rights of others or by consenting to the authority of some State-like entity. The right to punish is typically seen to follow from the right to protect oneself from aggression (see Otsuka, 2003, ch. 3; Quinn, 1994; Hurka, 2011, ch. 7). Since it is permissible to organize such defenses around your person as will deter anyone from infringing your rights by imposing high costs on any *attempt* at infringing your rights (e.g., setting barbed wires around your house to deter intruders), it is also permissible to protect yourself through measures that will deter anyone from infringing your rights by imposing high costs on anyone who *successfully* infringes your rights.

On a range of other views, while individuals have the right to defend themselves against aggression, the right to administer justice is solely the State's, which, if it is a just State, derives this right from the obligation of its citizens to obey just laws (Duus-Otterström and Kelly, 2019; Rawls, 1971, §51). Any just society must proscribe certain sorts of harmful behaviour – assault, harrassment, intimidation, etc. – which it must aim to prevent by what means are consistent with respect for the basic rights and liberties of all citizens. Punishment, restorative justice, and public safety institutions, if they are effective at depressing violent behaviour, compensating victims, healing community rifts, and if they are appropriately constrained by laws of due process, become means for enforcing the law that are consistent with respect for the rights of all (Bedau, 2004). Some retributivists (e.g., Duff, 2001) argue that punishment also has important value as an expression of official disapproval of criminal acts. Moreover, it is argued it would be unfair to law-abiding citizens not to enforce prohibitions against criminal behaviour, as this would create a class of harmful free-riders. On this picture, then, the state has a *duty* to use appropriate resources to enforce citizens' rights.

I am leaving it open what specific forms of criminal justice and public safety institutions are legitimate, e.g., whether prison sentences or violent institutions are permissible. All that matters for our purposes is that, on both these views, there exist permissible means for enforcing individual rights, permissible means, and that these permissible means, as an unfortunate matter of empirical fact, are imperfect and may occasionally lead to the misenforcement of rights. Indeed this is something which any plausible theory of rights should accept, but it meshes very poorly with EL.

Let us consider the Lockean picture first, and assume that individuals have the full range of enforcement rights, unless they voluntarily alienate them. It is a conspicuous fact about criminal justice that even the best methods available for identifying those guilty of crimes are imperfect. Moreover, our methods for determining what kind of punishment is justified, or what level of compensation it is right to extract, in light of the the severity of the crime, and of our interest in helping victims of crime and in deterring future crime, are also, of necessity, imperfect. Sometimes we get it right, sometimes we get it wrong. Even so, according to Nozick (1974, 107), provided the method we use for judging suspects and sentencing those found guilty is fair and is the most reliable method available, and provided we can demonstrate the fairness and reliability of our method to those upon whom we would use it, then we are permitted to render justice on those whom we find guilty by this method. Similar constraints presumably also justify public safety institutions, even if these institutions are also imperfect and carry the risk of rights-misenforcement (e.g., unintended harm to innocents in the course of responding to active shooter situations). And surely we may defend ourselves from aggression, even if we know there is a chance that in the heat of the moment we cause more harm in defending ourselves than is proportionate (e.g., I push my aggressor back, accidentally sending him careening out a window).

This view of the permissibility of rights-enforcement is plausible enough, but it contradicts EL. A first point to recognize here is that, while any individual has the right to defend themselves and seek redress from wrong-doers, they do not have a duty to do so. If you steal a loaf of bread from me, it may be more trouble than it is worth to me to get you to pay me back, or I may simply be forgiving and choose not the press the matter. If you raise your fist to hit me, I can just turn the other cheek. We might just choose to trust each other instead of establishing public safety institutions. A second point to underscore is that if we misenforce our rights by mistakenly punishing an innocent person for a crime or overcharging them for compensation, or by using excessive force in the course of defending ourselves or others, having used a method which we knew sometimes leads to such outcomes, then, however justified and blameless we are in having relied on such methods, we have infringed their rights. Indeed, it is precisely because unreliable processes of justice threaten regular infringements of the rights of the innocent that Nozick (1974, pp. 106-107) imposes epistemic constraints on their permissible use. But, to reiterate, even the most reliable enforcement methods are demonstrably imperfect, and therefore run a foreseeable risk of infringing the rights of the innocent. In contrast, one *is* guaranteed to infringe no one's rights if one never even attempts to defend anyone or to seek redress for wrongs. Thus if individuals face a choice between enforcing their rights through a highly reliable but imperfect procedure, or doing nothing, EL requires they do nothing.

Things don't change if we switch to the more Statist picture, or assume that individuals have voluntarily alienated their rights to private enforcement of their rights, in exchange for the protective services of some State-like entity. True, under such assumptions, the State has a duty to those subject to its power to protect their rights. But the State may no more than any of its citizens intentionally infringe the rights of some as a means of preventing the violation of the rights of others. After all, we generally regard states as having a greater obligation not to go to war and massacre civilians than to prevent other states that have gone to war from massacring a similar number of civilians. If the duty not to infringe anyone's rights is a perfect duty, as hardline deontologists maintain, then EL applies to the state, and it must sooner let its citizens suffer the greatest indignities than fail to guarantee that it itself does not infringe their rights. The state can guarantee this, but only provided it establishes no institutions endowed with the slightest power of coercion – not even such powers of coercion as are minimally necessary to ensure the state can respond to mass outbreaks of violence, penalize dangerous behaviour, or ensure the compensation of victims of crime. The right to self-defense remains unexercisable so long as there is a foreseeable chance that you might, in the heat of the moment, use excessive force to defend yourself.

Strict deontologists face a dilemma: either accept that EL is false, and that, under uncertainty, respect for rights does not imply an unlimited priority for liberty over axiological considerations, or accept that, under uncertainty, most rights are almost always unenforceable against those who would infringe them. Taking the second prong of this dilemma is extremely costly to deontology's credibility, because it completely devalues the very rights whose sanctity it insists upon. Who needs rights if you can't enforce them except (maybe) in the rarest of cases (e.g., a murderer is caught red-handed)? Everyone in society would recognize it as being in their interest to have rights that were slightly less uninfringeable. Risks of having their rights infringed can permissibly be imposed on individuals when countervailing moral considerations are sufficiently strong, if this means their rights could actually be enforced under the extremely common context of uncertainty. Indeed, for rights to adequately protect us against being instrumentalized in the service of maximizing the impartial good, it had better be that we can permissibly defend ourselves from attempts to so instrumentalize us. Moreover, such insistence on the inviolability of rights perversely punishes those who actually do respect the rights of others by incentivizing harmful free-riding on their efforts.

Taking the first prong of this dilemma is equally costly for absolutism, because if strong counterveiling axiological considerations can license inroads into protected liberties under the context of uncertainty, no grounds are left for resisting the moderate deontologist's view that strong counterveiling axiological considerations may license similar inroads into protected liberties under the context of certainty. It is of course open to the hard-line deontologist to retreat from absolutism under uncertainty while insisting upon absolutism under certainty. In my framework, this amounts to accepting L but rejecting EL. This is an available position in logical space, but it is not one that is easy to justify or motivate. The usual justifications for deontological constraints don't rely on the assumption of certainty. They appeal instead to ideas of respect or the inviolability of persons, and these don't suggest any natural asymmetries between different epistemic contexts. So it will go hard for absolutists if they attempt to justify an asymmetry by relying on these standard explanations.

For example, on Frances Kamm's (1995) influential view of rights, the possession of rights is identical with one's having a certain kind of moral status, namely, that of an inviolable person – a person that may not easily be instrumentalized in the pursuit of the greater good. According to Kamm (1996, p. 273), our inviolability stems from "certain properties [...] we have as individuals", and she hints that our most salient property is "having a rational will, whose consent we must seek before interfering with" (*ibid*, 276). The vocabulary of rights merely serves to spell out the content of this status, which it is good for us to have, for as Nagel (2007, p. 108) puts it: "What actually happens to us is not the only thing we care about: What *may* be done to us is also important, quite apart from whether or not it is done to us – and the same is true of what we may do as opposed to what we actually do." In other words, we are more valuable beings the more circumstances in which morality prohibits our being used in certain ways against our consent, and this is good for us because it demarcates us from other creatures whose interests may be more easily sacrificed, even if our unwillingness to kill innocents for the greater good means we are likelier to die as a result.

We are more inviolable the fewer exceptions are permitted to the principle that one may not harm innocents, and so to impute to individuals rights that admit of thresholds is to grant them less inviolability than to impute to them rights that admit of no thresholds. But by the same token we are more inviolable if morality prohibits the imposition of foreseeable and avoidable risks on us of having our rights infringed than if morality tolerates the imposition of such risks for the sake of other ends. If absolutists should try to ground the absolute character of side-constraints by appealing solely to the value of inviolability, then straightforwardly this argument must lead us to EL. Some additional consideration must be introduced to justify the exceptions that are made under uncertainty, such as the greater value for all of a system of rights that allows those rights to be enforced, or the lack of respect for rights-bearers that is expressed by a system of rights that doesn't allow the enforcement of those rights. But if the hard-liners can avail themselves of non-inviolability-based considerations to justify exceptions under uncertainty, why can we moderates not avail ourselves of quite similar considerations to justify exceptions under certainty? Johnson (2020), for example, argues that to always refuse to infringe the rights of one person to save innumerably more from certain death is disrespectful towards the innumerably many, because it betrays a moral outlook that refuses to take their interests into account. And Kamm herself emphasizes that the value of our saveability – the value we have as beings worth saving from harm – acts as a counterweight to the value of our inviolability.

Nozick's (1974, p. 31) own justification is Kantian: imputing rights to individuals just expresses and gives content to Kant's idea that we not treat people merely as means, but also always as ends in themselves. That is, individuals cannot be sacrificed for other ends – even very valuable ones, and even ones that are in their objective interests – without their consent. Alexander (2000), who defends absolutism, appears to accept this justification of constraints (see also O'Neill, 1980, p. 547). This view does not suggest interesting asymmetries between certainty and uncertainty; on the contrary, if respect for persons prohibits instrumentalizing them by harming them without their consent for the sake of some valuable end, why should it not also prohibit instrumentalizing the innocent by imposing foreseeable risks of unconsented-to rights infringements on everyone for the sake of some greatly valuable end? Considerations other than pure aversion to instrumentalization must be invoked to generate the desired asymmetry, which brings us back to where we were.

Vallentyne, Steiner, and Otsuka (2005) suggest a kind of reverse-engineering justification for rights, on which certain rights are to be imputed to individuals because the ownership of such rights best grounds and explains the intuitive wrongness of various forms of non-consensual interference with bodily integrity. There are structural constraints on what rights we can assign to individuals consistent with this explanatory aim (for example, all rights must be compossible; see Steiner, 1977), but if the ultimate justification for imputing certain rights to individuals is that this scheme of rights best rationalizes prior intuitions about what acts are permissible, then this story is not particularly favourable to the hard-liners, since moderates can lay claim to greater intuitive appeal for their system of rights.

Perhaps the absolutist deontologist might reply to all this that I have performed a sleight of hand, falsely portraying the reasons to reject EL as being the same kinds of reason as the ones I've advanced to reject absolutism under certainty. In rejecting EL, we weaken one class of rights to strengthen another, but in allowing rights to be infringed to avert bad consequences, we weaken rights to better pursue the Good. An absolutist could insist that rights may be traded for rights, but not for the impartial Good. Unfortunately, this response misses its mark, because in rejecting EL we are not trading rights for rights: EL does not imply that we *lack* enforcement rights – *de jure*, we have all the enforcement rights in the world, and these rights are absolute, insofar as it is impermissible (whatever the consequences) to prevent us from exercising them upon rights-violators – it simply implies that, *de facto*, these rights are never permissibly exercised. We therefore *gain* no rights by rejecting EL. Nor do we strengthen them, since they're already absolute. On the contrary, rejecting EL weakens all our rights; they just become more valuable in the process. Thus it *is* for axiological reasons that EL is implausible. The rub is that these reasons are hard to dismiss even for an absolutist.

In brief, once absolutism is rejected under the context of uncertainty, there is no longer any motivation for it under certainty. The reasons which justify limits on the scope of rights to constrain permissible action under uncertainty are the very same reasons which moderate deontologists appeal to to justify limits on the scope of rights under certainty. On pain of imposing arbitrary restrictions on morality, one must therefore be absolutist across the board, or moderate across the board. Of these two options, moderation is clearly to be preferred: it has greater intuitive appeal, it requires us to pay none of the pragmatic costs of EL, and it alone permits the enforcement of individuals' fundamental rights and liberties. Moreover, the moderate picture gives up no conceptual unity, since both the rights imputed to individuals and the limits imposed on those rights are grounded in the common value of respect for persons. With this in mind, I now propose some moderating amendments to the framework presented earlier.

7.3 The framework of liberty reconsidered

It is pretty clear intuitively what form moderate deontology must take. It must register the weight of both deontic reasons and axiological reasons, and allow for the balance of reasons to swing in either direction. There are many ways in which one can explicate this basic intuition, but I will take the simplest approach and model our all-things-considered moral reasons to act as some convex mixture of our deontic and axiological reasons.

To represent the demands of our axiological reasons on our choices, we can define for each individual k a value function $W_k : X \longrightarrow \mathbb{R}$, which encodes the demands of our favourite axiology on k. Thus $W_k(x) > W_k(y)$ implies that, from the perspective of k's axiological reasons alone, y is not permissibly chosen in any situation where x is an available alternative. I index W_k to the individual to make room for agent-relative permissions to weigh one's own interests and those of one's loved ones relatively more heavily (within limits) than those of others. Of course, your favourite axiology may make no room for partiality of this sort, in which case the indexing is unnecessary.

As we saw in the previous chapter, the weight of our deontic reasons can be represented by a function $\mathcal{D}_k(x) = \sum_{i=1}^{n} \beta_{xki} F_i(x)$, unique up to positive affine transformations, where $F_i(x)$ represents how severely individual i's rights are disrespected in x. Recall that β_{xki} is weighing parameter indexed to individual k, to allow k to weigh more heavily the rights of those who are closest to them and whom they can most easily save; the outcome x, to allow for rights which k is personally responsible for infringing by causing x to be weighed (much) more heavily, since our primary duty is to avoid violating rights, not prevent them from being violated; and to individual i, to ensure that, all else being equal, individuals whose rights are more disrespected get weightier β-parameters.

I propose that the ideal moderate deontological agent is rational and positively responsive to both their axiological and deontic reasons. This means that their choices respect the Savage axioms and that, all else being equal, they are reliably disposed to pursue better courses of action, and to avoid courses of action more disrespectful of the rights of others. Established results in social choice theory show that our ideal deontological agent can then be represented *as if* they are maximizing the probability weighted sum over every possible outcome of a convex mixture of their axiological and deontic reasons. So, in the Savage framework adopted in this paper, this means they can be represented as maximizing some function $V_k : \Gamma \to \mathbb{R}$ defined as follows:

$$V_k(f) = \sum_{l=1}^{m} P(s_l) \cdot [W_k(f(s_l) - \gamma D_k(f(s_l))],$$

where γ is a real constant reflecting the weight of deontic reasons relative to axiological reasons. Conversely, if an individual's choices can be represented as if they were maximizing some function V satisfying this definition, then the individual is rational and responds positively to their axiological and deontic reasons.

It is important to emphasize that V_k is not a decision procedure nor a description of anyone's moral psychology. V_k tells us what structure one's choices must have if they are to comply perfectly with the principles of moderate deontology, but says nothing substantive about what one should in fact choose (outside trivial cases), nor about what considerations should motivate us.

7.4 Assessing the revised framework

An attractive feature of this proposal is that it allows us to retain strong deontic commitments, while paying none of the pragmatic costs of absolutist deontology. Any individual whose choice pattern maximizes some V_k "reveals" a preference relation respecting the Savage axioms.[2] And this in turn means that a moderate deontologist will never choose dominated strategies, be money-pumpable, or be willing to pay an infinite price for certainty. At the same time, this framework is consistent with lending great weight to deontic reasons while still making room for the permissible enforcement of one's rights under uncertainty. Provided we are epistemically responsible, and reduce the chance of misenforcing our rights, countervailing evaluative reasons and reasons grounded in concern and respect for potential victims whose rights we could protect from intrusion may prove sufficiently weighty to override our deontic reasons not to infringe anyone's rights.

Moreover, this proposal allows us to defuse some of the standard objections to moderate deontological theories. I will now consider three that have been influential. First, Ellis (1992) and Christopher (2003) have argued that it is impossible to set the deontological threshold in a non-arbitrary fashion. Both argue that this is because deontic reasons and axiological reasons are incommensurable, in the sense that when they disagree their disagreement is either irresolvable, or it is resolved by one kind of reason always swamping the other. If disagreement is irresolvable, then trivially there is no discrete point at which the badness of consequences swamps the imperative not to infringe the rights of others, and likewise if one reason always swamps the other. Therefore, the argument goes, when moderate deontologists claim that you are permitted, say, to sacrifice the life of one innocent if by doing so you can save some minimal number N of lives, they must be pulling this number out a hat, as there can be no objective basis for their claim. But the difference between what is morally right and morally wrong cannot be a matter of arbitrary choice, therefore moderate deontology is false.

In response to this objection, note that if deontic and axiological reasons are incommensurable in Ellis and Alexander's sense, then it is impossible to be both rational and positively responsive to both axiological and deontic reasons. This is a strong and surprising implication of their claim, so they had better have a compelling argument. But their only gesture of support for it is the vague suggestion that "Deontologists treat killing an innocent person for others' benefit as intrinsically wrong. (...) On the other hand, for the consequentialist, consequences

[2] "Reveals" in the sense of revealed preference theory, meaning that their choices can be represented as if they had been induced by a binary betterness relation. The revealed preferences have no realist psychological interpretation. See Samuelson (1938), Sen (1971), and Thoma (n.d.)

are all that matter. The only thing that is intrinsically wrong for her is the act-type "failure to promote the best consequences"" (Alexander, 2000, p. 907). Ellis (1992, p. 862) adds, helpfully, that 'it seems plausible to think that these two sorts of consideration [intrinsic wrongness of acts and badness of consequences] will be incommensurable."

I'm not sure what to make of this. Perhaps what is being suggested is that, unlike axiological reasons, deontic reasons don't come in degrees, and therefore cannot be represented by a real-valued function, and for this reason continuous trade-offs between the two are impossible. But as I've argued in the previous chapter, deontic reasons *can* be represented by a real-valued function, D_k, so continuous trade-offs are possible. Indeed, even on absolutist pictures of rights, deontic reasons have degrees: your reasons not to steal my banana are less weighty than your reasons not to kill me. Perhaps Ellis and Alexander are simply incredulous at how the trade-offs could ever be carried out in practice. But we can lift the veil of mystery by constructing easy cases: because deontic and axiological reasons are both cardinally measurable, we can invent acts f and g such that our deontic reasons to choose f over g are arbitrarily weak (i.e. $\sum_{k=0}^{m} P(s_k)D_k(f(s_k)) - \sum_{k=0}^{m} P(s_k)D_k(g(s_k))$ is close to 0) and our axiological reasons to choose g over f are arbitrarily strong, or vice-versa. It strains credulity that by introducing an arbitrarily weak reason not to do something which we otherwise have arbitrarily strong reason to do, we thereby render the act impermissible. If this is right, though, then deontic and axiological are plainly commensurable.

With all this said, rationality and positive responsiveness alone don't tell us *where* the thresholds lie, i.e., how weighty both kinds of reasons are relative to one another. More substantive accounts of moderate deontology will have more to say about this, but one might still worry that any specific proposal of weights must have a conventional aspect. Fair enough. Our only guide to the "true" weights, I take it, is reflective equilibrium, hardly a precise measurement tool. If all we know is that it is wrong to sacrifice one innocent to save five, but permissible if to save 20, then any decision on *our* part to draw the line at 15, say, in the crafting of legal standards, will in a limited sense be arbitrary. But this is not a special problem for moderate deontology: egalitarians must weigh the interests of the worse off against the interests of the better off, absolutists must weigh weightier rights against less weighty rights, utilitarians must weigh better quality living against longer living, etc.[3] And we are all in the same boat regarding reflective equilibrium.

A second common objection to moderate deontology is that it leads to strange decisions in the neighbourhood of the thresholds. Suppose it is permissible to

[3] See also Zamir and Medina (2010, p. 46) for a point along this line.

torture an innocent to save n people from death, but not to save $n-1$; then, asks Alexander (2000, p. 900), supposing the police believe that by torturing some terrorist's mother they will induce the torturer to release his n^{th} captive, must the torture stop once the terrorist calls the police to let them know he's released one hostage, thereby reducing the number of people they could theoretically save from n to $n-1$? Alexander thinks a moderate deontologist must say it is permissible for the police to begin torturing the terrorist's mother, but require them to stop once the single hostage is released, which is very weird: terrorists shouldn't be able to "game" morality.

There is some major sleight of hand in this counterexample. Alexander starts by setting n as the minimal number such that it would be permissible to torture an innocent if we were *certain* to save n people by doing so, then presents a case where one has to decide whether to torture an innocent to try and save n innocents, but where we are initially quite *un*certain about how many we will actually save (clearly the police couldn't have been certain, since they turned out to be wrong), and then some later event occurs which *resolves* our uncertainty. But on the simple view that I've presented of how deontic and axiological reasons interact, if it is permissible to torture one innocent save at minimum n people with certainty, then it is not permissible to torture an innocent if you're far from certain to save n, and certain not to save any more than n. Many more lives would have to be at stake. To be sure, if you permissibly torture an innocent expecting to save a sufficient number N of lives, and it would be permissible to continue torturing them provided you could still expect to save N lives, but you later learn that the number of lives you can expect to save by continuing the torture is less than N, then continuing the torture is impermissible. But morality isn't being gamed here: *obviously* rationality may require you to alter your plans when you learn something which resolves your uncertainty about the world! I may find it worthwhile to walk across town to meet a date, but if, when I'm halfway there, she texts me to cancel, it would only compound my disappointment to keep walking to a cafe where I know I'll be alone.

A third and final classic objection to moderate deontology is that it capitulates to consequentialism. By admitting that people can sometimes be instrumentalized for the sake of averting terrible consequences, we're admitting that morality bottoms out in the assessment of the impartial goodness and badness of consequences. Indeed, the objection continues, it can be shown that moderate deontological theories are extensionally equivalent to the brand of consequentialism defended by Sen (1982) or Portmore (2001) which incorporates evaluator relativity. My own framework shows this, since any individual who conforms to the requirements of moderate deontology as I've laid them out can be represented as maximizing the value of some numerical function. And in this case, the difference between moderate deontology and consequentialism comes down to an indifferent choice of vocabulary, not a substantial disagreement about what morality is about and how one should think about one's decisions.

On the face of it, the complaint that deontology capitulates to consequentialism by making space for evaluative considerations seems a bit strange. One might as well say that consequentialism capitulates to deontology by making space for deontic considerations and permissible partiality. By admitting that the aim of doing good is habitually subordinated to considerations of right, and that anyway the good is to be evaluated from the position of the agent, it accepts that morality does not bottom out in the assessment of the impartial goodness and badness of consequences, but in individual agency.

More seriously, though, as Tennenbaum (2014) argues in response to Portmore, to argue that some theory is just a form of consequentialism it is not enough to show that the choices it induces could be represented *as if* it were maximizing the value of some numerical function. Any choice pattern can be so represented, if we are sufficiently creative in how we describe the objects of choice and throw in enough "mathematical fireworks" (Tennenbaum, 2017, p. 686). Consequentialism requires that the ordering "revealed" by our choices be determined by the social welfare function, not the other way around (Sen, 1986; J. A. Weymark, 1991; Ponthière, 2003), otherwise it is no moral theory, only a trifling mathematical thesis regarding the representation properties of certain choice functions.

Moderate deontology is a form of deontology because it preserves what is distinctive about deontology generally, namely the commitments to the priority of rights over value, and to the agent-relativity of certain reasons. As Sen (1982, p. 22) highlights, there are three ways in which reasons might be agent-relative in any given case. Self-evaluation relativity: there may be reasons for me to do things which are not reasons for you to do that same thing, e.g., that f would benefit my friends and g would benefit yours give me a reason to do f, but not you. Doer relativity: there may be reasons which prohibit me from doing something which however do not require me to prevent you from doing it, e.g., as Nagel (1980, p. 119) argues, it may be that I ought not to twist a child's arm to bring about some good consequence even though it would not be required of me to forego a comparable benefit to prevent someone else from twisting a child's arm. And viewer relativity: I may be permitted to do something which you are required to prevent, e.g., suppose two people need the same abandoned boat to save their own children; then although either is permitted to take the boat, both are required by the balance of their own evaluative reasons to prevent the other from taking it. My framework preserves all three types of agent-relativity. True, the side-constraints rights place on permissible action are spongy, but as Nagel (2007, p. 106) rightly notes: this does "not change the basic character of the right, since the threshold will be high enough so that the impermissibility of torture or murder to prevent evils below it cannot be explained in terms of the agent-neutral badness of torture or murder alone."

In contrast, agent-relative consequentialism does lose what is distinctive about consequentialism generally, namely the commitment to all moral reasons being agent-neutral. As Brown (2011, p. 759) has argued, a necessary condition for a theory to qualify as consequentialist is that it instructs every single individual to maximize exactly one common and complete ranking relation on the set of possible outcomes. This characterization is quite minimal, but it implies that consequentialism at a minimum violates viewer relativity, and in fact most consequentialist theories violate all three forms of relativity. Indeed, it is for this very reason that Sen himself insists that his own preferred theory, which does make room for all three kinds of relativity, is not a brand of consequentialism, but rather a third way between consequentialism and deontology which he alternately calls a "goal-rights system" or "consequence-based evaluation." Given how far removed Sen's view actually is from consequentialism, it's not so surprising that moderate deontology should turn out to be extensionally equivalent to it. But this just shows that worries of our having capitulated to the cold utilitarian calculus of consequences are misplaced, and anyway one might reasonably take the view that agreement with Amartya Sen is a thing to be celebrated.

Summary

Rights are not worth the price of purchase if they can't be enforced. When we try to extend absolutist commitments from the context of certainty to the context of uncertainty, we find we can do so only by paying important pragmatic costs, and by *de facto* denying everyone the permission to enforce anyone's rights. Creating permissions for the enforcement of individual rights therefore requires abandoning absolutism under uncertainty, which undermines the motivation for absolutism more generally, and pushes us towards a more moderate picture of deontology. I have tried to state this view precisely, and to show that it is in fact a very attractive picture of morality: it complies with all the norms of rationality, gives rights their full value and carves robust boundaries around personal liberty while making appropriate room for axiological concerns, and retaining the distinctive commitments of deontology. Moreover the standard objections to this sort of view are unconvincing, as they largely rest on a misunderstanding of moderate deontology's structure.

****

To conclude this long study on freedom, I wish to offer some parting thoughts on how I see the concern for the opportunity aspect of freedom interacting with the concern for the process aspect of freedom, to help make sense of how everything that has been argued so far hangs together.

My view is that the concern for the opportunity aspect of freedom should be folded into our assessments of how well individuals' lives are going, while our concern for the process aspect of freedom should be folded into our ethics of respect. Or in other words: in its opportunity aspect, freedom is a component of well-being, i.e., an argument in the value function W_k which captures our welfarist concerns, whereas in its process aspect, freedom is the grounds for certain kinds of side-constraints on permissible action, i.e., an argument in the value function D_k which represents our deontological concerns. An attractive feature of this view is that it unifies the disparate values that have traditionally animated the debates between republicans, liberals, and capability theorists. W_k registers the importance of making individuals more capable, of giving them more valuable opportunities, and of immunizing them against domination, while D_k registers the special importance of non-interference in protected spheres of choice.

Naturally, our welfarist and deontological concerns are not exhausted by our concerns for freedom: we may care about how well people's lives *actually* turn out, not just how good their opportunities are, and we may think that respect for others requires not just that we refrain from interfering with them, but also that we not lie to them or think of them unkindly or intend them harm. Still, much of morality hangs just on properly appreciating the value of freedom.o

Appendix A

Measuring distance across possible worlds

We start by defining a binary, nonsymmetric and transitive relation $\trianglerighteq$ on the set $\mathcal{W} = \{w_1, w_2, ...\}$ of possible worlds, interpreted as the "at least as dissimilar from the actual world as" relation. I define $\triangleright$ and $\bowtie$ as the "more dissimilar than" and "equally similar to" parts, respectively, of this relation. Because some counterfactuals are true, and true counterfactuals are validated by facts about which possible world is most similar to the actual world, it follows that there are facts about which possible worlds are more similar than others to the actual world – i.e. we know $\trianglerighteq$ is nonempty. And not only are there such facts, but they are epistemically accessible too, since they can be discovered by investigating counterfactuals. If any two possible worlds differ in some respect, there must be a true counterfactual which reveals this difference. Therefore, a large enough body of true counterfactuals determines a complete ordering of $\mathcal{W}$. We then define the world-distance function $d : \mathcal{W} \longrightarrow \mathbb{R}$ which we require to satisfy the property that $d(w_1) \geq d(w_2) \iff w_1 \trianglerighteq w_2$.

This doesn't yet bring us to shore, as $d(.)$ only captures ordinal information about world similarity, and I promised a cardinal measure. But notice that just as we can ask which worlds are most similar to the actual world, we can also ask which worlds are most similar to each other. These facts are likewise discovered by investigating counterfactuals: on Lewis's analysis, the counterfactual $A\Box \longrightarrow (B\Box \longrightarrow C)$ is true if and only if there is a world w in which A was the case in which the embedded counterfactual $B\Box \longrightarrow C$ is true which is closer to the actual world than any world in which A was the case but the embedded counterfactual was false, and for that counterfactual to be true in w is just for it to be the case that there is a world w' in which B and C were the case which is closer to w than any world in which B was the case but not C.

And this gives us a way of constructing a quaternary preorder $\trianglerighteq_d$ on $\mathcal{W} \times \mathcal{W}$, which ranks pairs of worlds by their differences in similarity from the actual world.

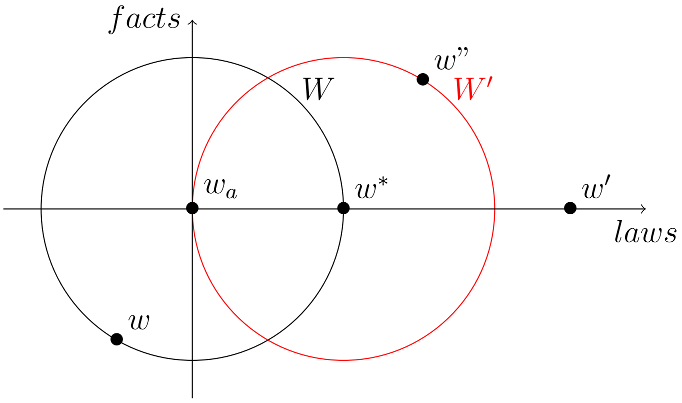
The procedure I am about to describe is represented graphically below in figure 1. Suppose w is more similar to the actual world w_a than w' is. Let W be the set of all possible worlds that are equally similar from w_a as w is, and let w^* be the world in W which is most similar to w'. If w^* is more similar to w_a than to w', then the difference in similarity from w_a between w and w_a is less than the difference in similarity from the actual world between w^* and w'. This is noted $(w^*, w') \rhd_d (w, w_a)$. Now suppose instead we want to find worlds $w, w', w”$ such that $(w, w_a) \bowtie_d (w^*, w”)$. This is easily done by taking the set of all possible worlds W' that are equally dissimilar from w^* as w_a is, then picking some arbitrary element $w”$ of this set. Continuing in like manner, we form a rich ordering. We now require that $d(.)$ satisfy the additional property that $d(w_1) - d(w_2) \geq d(w_3) - d(w_4) \iff (w_1, w_2) \trianglerighteq_d (w_3, w_4)$.

This and the previous requirement I imposed on $d(.)$ are very natural restrictions on a distance function, but there are alternatives. For example, we could set $\overline{d}(w) = ln(d(w))$. This distance function would continue to satisfy the requirement that $\overline{d}(w) \geq \overline{d}(w') \iff w \trianglerighteq w'$, so it is an order-preserving transformation of $d(.)$, but it makes distance multiplicative in the dissimilarity of worlds from the actual world rather than additive, so it does not preserve intervals. This would be a proper measure, but not a natural one. For an analogy, consider the measurement of length. We normally require the overall length of two objects to be equal to be the sum of their lengths, this numerical convention corresponding to the empirical operation of concatenating two rods by abutting them end to end. But we could instead require the overall length of two objects to be equal to the square root of the sum of their squares, this convention intuitively corresponding to the empirical operation of concatenating two rods by abutting them at a right angle. We don't do this though, because it's unintuitive and mathematically cumbersome, and similar considerations apply here. See Weymark (2005) for more.

And with this last requirement, we guarantee the uniqueness of $d(.)$ up to positive affine transformations. This is easy to verify: take any worlds w_1, w_2, w_3 such that $d(w_1) > d(w_2) > d(w_3)$ and $d(w_1) - d(w_2) = d(w_2) - d(w_3)$; given the two requirements I've imposed on $d(.)$, any admissible transformation of $d(.)$ must preserve the truth of these two expressions. But crucially, any two triples satisfying these expressions must be related to one another by a positive affine transformation (Ellsberg 1952, 534).

A graphical representation may help to visualize things. Suppose we represent the set of possible worlds as the 2-dimensional (because we're supposing that world-similarity varies along 2 dimensions, noted here “$laws$” and “$facts$”) euclidean space E^2. Points on the euclidean plane are possible worlds, and their distance from the origin represents their distance from the actual world, which is located at the origin. In figure 1 below, W is the circle centered on w_a with the radius of length $w_a w$, and w^* is the point where the line (ow') intersects with W. (Note that W doesn't need to be represented by a circle: any isodistance contour would do. We could replace the curves in this figure with straight lines, or with sawtooth edges, or elliptical curves. This is important, because otherwise the reader might think I was committed to a Euclidean measure of distance, whereas it is in no way implied by my method of construction that world-distance must be Euclidean.) If w^* is closer to o than to w', then $ow^* < w^*w'$. Since $ow^* = ow$, it follows that the difference in similarity from o between o and w is smaller than the difference in similarity from o of w^* and w'. Likewise, W' is the circle centered on w^* with radius w^*w_a and w" is any point on this circle. The equality in the difference in similarity between w_a and w and the difference in similarity between w" and w^* shows up graphically in the fact that $w_a w$ and w^*w" are two radii of circles with the same radius.

Figure 1



Some final remarks on this whole construction procedure are in order. It's important to realize that this entire construction only makes sense if we assume, first, that the set of possible worlds is densely ordered by similarity from the actual world (i.e. for any two worlds $w, w' : w \triangleright w'$, there is a world w" $: w \triangleright w$" $\triangleright w'$), and second, that the set of possible worlds has the least upper-bound property (i.e. for any set S of possible worlds such that there is a world w which is at least as dissimilar from w_a as any element of S, there is a world w' which is at least as dissimilar from w_a as any element of S such that any world more similar to w_a than w' is more similar to w_a than at least one element of S). I have no argument for these assumptions. On the face of it, though, both strike me as plausible. And in fact, the following argument by Lewis against the so-called limit assumption suggests that at least *some of the time* both conditions are satisfied.

Suppose we entertain the counterfactual supposition that at this point

> there appears a line more than one inch long. (Actually it is just under an inch.) ...But how long is the line in the closest worlds with a line more than an inch long? If it is 1 + x” for any x however small, why are there not other worlds still closer to ours in which it is 1 + 1/2x”, a length still closer to its actual length? ... Just as there is no shortest possible length above 1”, so there is no closest world to ours among the worlds with lines more than an inch long. (Lewis 1973, pp 20-21)

In this example, possible worlds are being ordered (literally!) by the real line, which is densely ordered and has the least upper bound property. So if Lewis's rather persuasive reasoning in this instance is sound, then at least some subsets of $\mathcal{W}$ are densely ordered and satisfy the least upper bound property. This is far from a demonstration that the properties hold for $\mathcal{W}$ as a whole, but at least as a default position I feel we are justified in assuming this. The onus is now on skeptics to show that the subset of worlds implicated by Lewis's example is exceptional.

Appendix B

Proof of Theorem 4.2.2

Necessity is obvious, so I only prove the sufficiency part of the theorem. Let $\succsim$ satisfy (INS), (SM), (MM) and (S). I first show that

> **Indifference**: $A \sim B$ for all $A, B \in Z$ such that $max\{1, |A \cap \mathcal{M}|\} = max\{1, |B \cap \mathcal{M}|\}$ and $|A| = |B|$.

Proof. We use an inductive argument to show this. So first I prove

> (Inductive step) For every positive integer n, if $[max\{1, |A' \cap \mathcal{M}|\} = max\{1, |B' \cap \mathcal{M}|\} = n$ and $|A'| = |B'|] \Longrightarrow (A' \sim B')$ for all $A', B' \in Z$, then $[max\{1, |A \cap \mathcal{M}|\} = max\{1, |B \cap \mathcal{M}|\} = n+1$ and $|A| = |B|] \Longrightarrow (A \sim B)$.

Consider a positive integer n such that $[max\{1, |A' \cap \mathcal{M}|\} = max\{1, |B' \cap \mathcal{M}|\} = n$ and $|A'| = |B'|] \Longrightarrow (A' \sim B')$ for all $A', B' \in Z$. Let $max\{1, |A \cap \mathcal{M}|\} = max\{1, |B \cap \mathcal{M}|\} = n+1$ and $|A| = |B|$. Let $C \subsetneq A$ such that $max\{1, |C \cap \mathcal{M}|\} = n$ and $A \setminus C = \{x\}$. Clearly, $x \in \mathcal{M}$, otherwise it would follow that $A \cap \mathcal{M} = C \cap \mathcal{M}$ and therefore that $max\{1, |A \cap \mathcal{M}|\} = max\{1, |C \cap \mathcal{M}|\}$, contradicting our assumption. Now, either

> (a) $x \in B$,

or

> (b) $x \notin B$

Suppose (a) holds. Then let $D = B \setminus \{x\}$. Because $x \in \mathcal{M}$ and $max\{1, |B \cap \mathcal{M}|\} = n+1 \geq 2$, it follows that $max\{1, |D \cap \mathcal{M}|\} = n$. Thus $max\{1, |C \cap \mathcal{M}|\} = max\{1, |D \cap \mathcal{M}|\} = n$. Further, $|C| = |D|$, since $|C \cup \{x\}| = |A| = |B| = |D \cup \{x\}|$. Given our inductive assumption, this implies $C \sim D$. (S) then implies $C \cup \{x\} \sim D \cup \{x\}$, i.e. $A \sim B$.

Now suppose (b) holds. There must be some $y \in B$ such that (i) $y \notin A$ and (ii) $max\{1, |B \setminus \{y\}|\} = n$. Otherwise, if all $y \in B$ for which (ii) is true also belong to A, then the union E of all such y is a subset of $A \setminus \{x\}$. But clearly, $max\{1, |E \cap \mathcal{M}|\} = n + 1$, thus $max\{1, |E \cup \{x\}|\} = n + 2$, which contradicts our assumption that $max\{1, |A \cap \mathcal{M}|\} = n + 1$ because $E \cup \{x\} \subset A$.

With this in mind, for some y satisfying (i) and (ii) let $F = B \setminus \{y\}$. Then $max\{1, |C \cap \mathcal{M}|\} = max\{1, |F \cap \mathcal{M}|\} = n$, and moreover $|C| = |F|$, since $|C \cup \{x\}| = |A| = |B| = |F \cup \{y\}|$. Thus, by assumption, $C \sim F$, and hence by (S) it follows that $(C \cup \{x\}) \sim (F \cup \{x\})$, i.e.

(c) $A \sim (F \cup \{x\})$.

Now consider B and $F \cup \{x\}$. Since $max\{1, |(F \cup \{x\}) \cap \mathcal{M}|\} = max\{1, |B \cap \mathcal{M}|\} = n+1$, it follows that $|B \cap \mathcal{M}| \geq 2$, and therefore that $|F \cap \mathcal{M}| = |B \cap \mathcal{M}| - 1 \geq 1$. Let $z \in (F \cap \mathcal{M}) \subset B$. Noting that $max\{1, |(F \cup \{x\}) \setminus \{z\}|\} = max\{1, |B \setminus \{z\}|\} = n$ and $|(F \cup \{x\}) \setminus \{z\}| = |B \setminus \{z\}|$, we have $((F \cup \{x\}) \setminus \{z\}) \sim (B \setminus \{z\})$. Hence, by (S), we have

(d) $(F \cup \{x\}) \sim B$.

From (c) and (d) and the transitivity of $\succsim$, we obtain $A \sim B$.

This completes the proof of the inductive step. I now prove

(inductive premise) $\forall A, B \in Z$ such that $max\{1, |A \cap \mathcal{M}|\} = max\{1, |B \cap \mathcal{M}|\} = 1$, and $|A| = |B|$, $A \sim B$

Obviously, for all singleton sets $A \in Z$, $max\{1, |A \cap \mathcal{M}|\} = 1$, and we know from (INS) that for any two singleton sets $A, B \in Z$, $A \sim B$. Furthermore, we know that for all non-singleton sets $A \in Z$ such that $max\{1, |A \cap \mathcal{M}|\} = 1$, at most one element in A can belong to $\mathcal{M}$, otherwise $|A \cap \mathcal{M}| > 1$, and therefore $max\{1, |A \cap \mathcal{M}|\} > 1$, contradicting our assumption. So, let $A = \{a_1, ..., a_m\}, B = \{b_1, ..., b_m\}$ be any two sets of equal size containing at most one element that also belongs to $\mathcal{M}$. If there is an element in A that also belongs to $\mathcal{M}$, let this unique element be denoted a_1; otherwise, let a_1 denote any arbitrary element from A. Let b_1 be the element from B likewise defined. From (INS), $\{a_1\} \sim \{b_2\}$, and since we know that all remaining a_j, b_j do not belong to $\mathcal{M}$, we know from (S) that $\{a_1\} \cup \{a_2\} \sim \{b_2\} \cup \{b_2\}$, and so repeated applications of (S) yield $A \sim B$.

This establishes the inductive premise. □

Having established the indifference condition, I now prove:

Dominance: $A \succ B$ for all $A, B \in Z$ such that either:

$max\{1, |A \cap \mathcal{M}|\} > max\{1, |B \cap \mathcal{M}|\}$, or

$max\{1, |A \cap \mathcal{M}|\} = |max\{1, |B \cap \mathcal{M}|\}$ and $|A| > |B|$.

Proof. Consider $A, B \in Z$ such that $max\{1, |A \cap \mathcal{M}|\} > max\{1, |B \cap \mathcal{M}|\}$. Necessarily, $|A \cap \mathcal{M}| > |B \cap \mathcal{B}|$ and $|A \cap \mathcal{M}| > 1$, so let $a_1, ..., a_n$ denote the n elements of $A \cap \mathcal{M}$. If $B \cap \mathcal{M} \neq \varnothing$, then let $b_1, ..., b_m$ denote the $m < n$ elements of $B \cap \mathcal{M}$.

Let $H = \{x \in A | x \notin \mathcal{M}\}, G = \{x \in B | x \notin \mathcal{M}\}$. From (MM), we know $H \cup \{a_1, a_2\} \succ G \cup \{b_1\}$. If there remain no more $b_i \in B \cap \mathcal{M}$, this means $H \cup \{a_1, a_2\} \succ B$. If there do remain elements $b_i \in B \cap \mathcal{M}$, then by repeated applications of (S) we get $H \cup \{a_1, ..., a_{m+1}\} \succ G \cup \{b_1, ..., b_m\}$, i.e. $H \cup \{a_1, ..., a_{m+1}\} \succ B$. Either way then, $H \cup \{a_1, ..., a_{m+1}\} \succ B$, and if $n = m + 1$, this just means $A \succ B$. If $n > m + 1$, then, since (MM) implies $H \cup \{a_1, a_2\} \succ H \cup \{a_1\}$, successive applications of (S) give us $H \cup \{a_1, ..., a_n\} \succ H \cup \{a_1, ..., a_{n-1}\}$, i.e. $A \succ \{a_1, ..., a_{n-1}\}$. By the transitivity of $\succsim$ it then follows that $A \succ B$. Thus, for all A, B, $max\{1, |A \cap \mathcal{M}|\} > max\{1, |B \cap \mathcal{M}|\}$ implies $A \succ B$.

Now consider A, B such that $max\{1, |A \cap \mathcal{M}|\} = |max\{1, |B \cap \mathcal{M}|\}$ and $|A| > |B|$. Let $G \subset A$ be such that $max\{1, |G \cap \mathcal{M}|\} = |max\{1, |B \cap \mathcal{M}|\}$ and $|G| = |B|$, and let $H = A \setminus G$. Further, let $g_1, ..., g_n$ and $b_1, ..., b_n$ denote the n elements of G and B, respectively, and let $h_1, ..., h_m$ denote the m elements of H. Further if G (and therefore B) contains $k > 0$ minimally valuable elements, then, for all $1 \leq i \leq k$, let a_i,b_i denote some minimally valuable element belonging to A, B. Given how I've now ordered the elements of A and B, it follows that, for all a_i, b_i, $a_i \in \mathcal{M} \iff b_i \in \mathcal{M}$.

(INS) implies $\{a_1\} \sim \{b_1$, and (SM) implies $\{h_1, a_1\} \succ \{a_1\}$. Successive applications of (S) imply $H \cup \{a_1\} \succ \{h_1, ..., h_{m-1}, a_1$. By the transitivity of $\succsim$, it follows that $H \cup \{a_1\} \succ \{b_1\}$. Successive applications of (S) now imply $H \cup \{a_1, ..., a_n\} \succ \{b_1, ..., b_n\}$, i.e. $A \succ B$. Thus, for all A, B such that $max\{1, |A \cap \mathcal{M}|\} = |max\{1, |B \cap \mathcal{M}|\}$ and $|A| > |B|$ implies $A \succ B$. □

Given the indifference condition, the dominance condition completes the proof.

Appendix C

Proof that D is monotonic, dispersing, continuous, and translation invariant

1. I first show that D satisfies monotonicity. Let A, B be arbitrary sets such that $A \subset B$. By definition, for any $b_i \in B \setminus A$, the function g_{b_i} reaches its maximum value of 1 at $x = d(b_i)$, while for every other $a_i \in A$, g_{a_i} has a value of strictly less than 1 at $x = d(b)$. Therefore there is a collection of open intervals in $\mathbb{R}^n$ such that $max(g_{a_1}, ..., g_{a_n}, g_{b_1}, ..., g_{b_m})$ is greater over those intervals than $max(g_{a_1}, ..., g_{a_n})$. These open intervals will be the neighbourhood of points around each point $d(b_i)$ that are strictly closer to $d(b_i)$ than any other element of B, which are obviously open intervals since B is a finite set of points. Further, there is no point in $\mathbb{R}^n$ over which $max(g_{a_1}, ..., g_{a_n}, g_{b_1}, ..., g_{b_m})$ is strictly lesser than $max(g_{a_1}, ..., g_{a_n})$. Therefore $D(B) > D(A)$.

2. Showing that D satisfies dispersion is harder work. Let's start by considering a set A of points $d(a_1), ..., d(a_m) \in \mathbb{R}^n$. Let $\mathcal{K}$ be the convex hull of A and assume that $d(a_1)$ does not lie in the interior of $\mathcal{K}$. Let $\mathcal{H} \subset \mathbb{R}^n$ be a hyperplane which contains $d(a_1)$ and which does not intersect the interior of $\mathcal{K}$. Let $\mathbf{v}$ be a unit length vector orthogonal to $\mathcal{H}$.

Let us now "pull" $d(a_1)$ away from $\mathcal{K}$ while keeping $d(a_2), ..., d(a_m)$ fixed. More precisely, for $s \geq 0$, we consider the family $d(a_1, s) := d(a_1) + s\mathbf{v}$. We let $g_s(x) = max_i(e^{-|x-d(a_i,s)|^2})$ and we let $D(s) = \int_{\mathbb{R}} g_s$. Note that g_s is simply an abbreviated notation for the integrand of my diversity function D, and accordingly $D(s)$ denotes the diversity of the set of options $a_1, ..., a_m$ as we vary the position of $d(a_1)$. What we want to show is that as we "pull" $d(a_1)$ away from all the other points – i.e. as we make a_1 increasingly dissimilar from every other alternative in A – the diversity of A increases.

More precisely, we wish to prove that $D(s)$ is nondecreasing. To prove this, we begin by taking an arbitrary point $\mathbf{b} \in \mathcal{H}$. We consider the line $l(t) = t\mathbf{v} + \mathbf{b}$. We will now study the restriction of g_s to the line l.

By applying an affine transformation and choosing an appropriate orthogonal basis we may as well assume that $\mathbf{v} = e_1$ and that $\mathcal{H} = \text{span}\{e_2, ..., e_n\}$. Then $g_{s|l} = max_{i=1}^{m}(C_i e^{-(t-(d(a_i))_1(s))^2})$, where $C_i = e^{-(d(a_i)_2)^2} ... e^{-(d(a_i)_n)^2}$. Note that $(d(a_1)_1(s) = s$ while for $j = 2, ..., m$, we have $(d(a_j))_1(s) = (d(a_i))_1$.

Lemma C.0.1. *If $d(a_i)_1 \leq 0$ for $i = 2, ..., m$, then $\int_l g_{s|l}$ is nondecreasing in s, for $s \geq 0$.*

Proof. By translation, we may as well assume $s = 0$.

It's enough to show that the set of $t \in \mathbb{R}$ on which e^{-t^2} is greater or equal to $Ce^{-(t-\alpha)^2}$ forms an interval of the form $[d, \infty)$. – if $\alpha = 0$, then this is clear. So assume $\alpha < 0$.

We consider the quotient $e^{-t^2}/(Ce^{-(t-\alpha)^2}) = e^{a^{2-2t\alpha}}/C$. We ask when this is ≥ 1. Taking logs, this is the same as asking when $-logC + (-2t\alpha + \alpha^2)$ is ≥ 0. i.e. $t \geq (-\alpha^2 + logC)/2|\alpha|)$. The right-hand term of this inequality is a real number d, thus $t \in [d, \infty)$.

□

Corollary C.0.1.1. *$D(s) = \int_{\mathbb{R}^{n-1}}(\int_{l_{\mathbf{b}}} g(s)|l_{\mathbf{b}})$ is nondecreasing.*

Proof. Since the integrand is nondecreasing, the integral is nondecreasing. □

3. D is trivially continuous. Gaussian functions are continuous, the max function is continuous, and the integral function is continuous. Any function which can be decomposed into continuous functions of other continuous functions is itself continuous.

4. Let A, B be two equinumerous sets of objects such that $d(b_1), ..., d(b_m)$ can be obtained from $d(a_1), ..., d(a_m)$ by translation by some vector c. Then $D(B) = \int_{\mathbb{R}^n} max(g_{a_1+c_1}, ..., g_{a_m+c_m})$ which is obviously equal to $D(A) = \int_{\mathbb{R}^n} max(g_{a_1}, ..., g_{a_m})$, since the area under a curve does not change when you translate every point in the curve by a common vector.

Appendix D

Formal example of a liberal and illiberal state

Consider two possible worlds x, y and let $I_x = I_y = \{1, 2\}$ be our population. $\Omega = \{\varnothing, a, \neg a, b, \neg b, a \wedge b, a \wedge \neg b, \neg a \wedge b, \neg a \wedge \neg b, a \vee b, a \vee \neg b, \neg a \vee b, \neg a \vee \neg b, a \vee \neg a\}$ is our Boolean algebra of propositions. Let's interpret a as "prisoner 1 cooperates" and b as "prisonner 2 cooperates."

Let $R_1 = \{\{a, \neg a\}\}$ and $R_2 = \{\{b, \neg b\}\}$.

$A^x_{11} = \{a, \neg a\}, A^x_{21} = \{b, \neg b\}$; $A^y_{11} = \{a\}, A^x_{21} = \{b\}$.

Therefore $A_x = ((\{a, \neg a\}), (\{b, \neg b\})), A_y = ((\{a\}), (\{b\}))$.

By my definitions, x is a liberal outcome – corresponding under my interpretation to the classic prisoners' dilemma where both players are free to choose whether to cooperate or defect – and y is an illiberal outcome – corresponding to a situation where both players are forced to cooperate with each other.

Appendix E

Measuring disrespect for rights

To construct the family of f_i measuring the overall extent to which an individual's rights have been disrespected, we first must have a way of measuring the extent to which any individual's rights have been violated, and a way of measuring the relative importance of these rights. Different ways of combining these measures will give us different families of f_i. Below, I show how to construct one particular family of f_i, but I will flag along the way where one could have made different choices.

Translated back into my framework's vocabulary, to say that some right R_{ij} has been infringed more-or-less in some outcome x is to say that more-or-fewer elements of R_{ij} are missing from A^x_{ij}. Some of the propositions you've a right to make true or have a claim to being true are *missing* from the set of propositions you can in fact make true/ have a claim to being true and are in fact true. Set out in this way, there is a natural upper and lower bound on how deeply any right may be infringed. The upper bound is reached when you can make true *none* of the propositions you've a right to (or when none of the propositions you've a claim to being true are in fact true) and the lower bound is reached when your right is not infringed at all. So let $M^x_{ij} = R_{ij} \setminus A$ So define $d^x_{ij} \in \mathbb{R}^+$, representing how deeply R_{ij} was infringed. Set $d^x_{ij} = 1 \iff M^x_{ij} = R_{ij}$, $d^x_{ij} = 0 \iff M^x_{ij} = \varnothing$, and make d^x_{ij} linear in the ratio of $|M^x_{ij}|$ to $|R_{ij}|$. Thus $d^x_{ij} = |M^x_{ij}|/|R_{ij}|$.

To represent the fact that some rights are more important than others, we need to work a bit harder for our meal. Let's assume, along with all theories of rights, that we are all born with the same natural rights. To the extent that our present rights differ from those we were born with, that is because since birth we have gained new bundles of rights, lost some bundles, or the content of some of our bundles of rights has changed (e.g. you voluntarily surrender some, but not all, of your liberties of movement – which means, actually, that one of youe original R_{ij} has been "replaced" by different R_{ij} that share some elements of the old). Let's further assume that everyone's rights matter equally at birth, something a plausible theory of rights must imply. Translated back into my framework, this means we can define for each individual a "natural" schedule of rights $\overline{R}_i = \{\overline{R}_{i1}, ..., \overline{R}_{im}\}$. Intuitively, if $\overline{R}_{11}$ is the bundle of rights which represents individual 1's natural right to free speech, then $\overline{R}_{21}$ is the bundle of rights which represents individual 2's natural right to free speech, and so on for every $j(1 \leq j \leq m)$.

Next, define for each i the set P_i of probability distributions P_i over $\overline{R}_i \cup \{\varnothing\}$ (or, more accurately, over a σ-algebra whose universe is $\overline{R}_i \cup \{\varnothing\}$), and a relation $\trianglerighteq_i$ over P_i ranking i's natural rights in importance. Different interpretations of $\trianglerighteq_i$ are possible. You may interpret it as a representation of the individuals' own preferences over their rights; this would be attractively liberal as it would sensitize evaluations of how disrespectful it is to violate your rights to your own evaluations of how important those rights are, and it allows different individuals to prioritize as they please their own rights. Of course, this interpretation raises worries about adaptive preferences, and that individuals might come to disregard certain rights because they have been induced to by people who do not want them to value those rights. In this case we wouldn't want our evaluations to be too sensitive to individuals' own evaluations. Regimes which brainwash their citizens into devaluing their rights should be penalized, not rewarded. So perhaps it is best to impose some objective constraints on the structure of $\trianglerighteq_i$, and let individuals fill in the remaining gaps. In this case $\trianglerighteq_i$ is interpreted as the preorder which is nearest to the individual's own preferences but which also satisfy the objective constraints imposed by the true theory of rights. I myself, being worried about adaptive preferences, favour this interpretations, though I would endorse relatively few constraints.

Now, Intuitively, the probability distribution $0.8R_1 + 0.2\{\varnothing\}$ represents a lottery that gives you a 0.8 chance of "losing" every right except R_1 (meaning it would be permissible to interfere with your doing what those rights normally entitle you to do) and a 0.2 chance of losing every right, period. Assume $\trianglerighteq_i$ satisfies the vNM axioms of expected utility theory, there is therefore a real-valued function $\overline{u}_i$, unique up to positive affine transformations, representing $\trianglerighteq$. Because R_i is finite and $\succsim_i$ is complete, there is a (not necessarily unique) R_{max} such that $R_j \succsim R_{j'}$ for all j', and I assume that all R_j strictly dominate the empty set. I then use the zero-one rule to normalize $\overline{u}_i(\{\varnothing\}) = 0$ and $\overline{u}_i(R_{max}) = 1$. This amounts to assuming that, at birth, your most important right is as important as my most important right, and it would be as bad if you were deprived of all rights as if I were deprived of all rights. Again, any plausible theory of rights implies this.

Finally, we use the function $\overline{u}_i$ representing the value of each individual's rights at birth to set the origin and unit of the function u_i which represents the value of their present rights. If individual i's schedule of rights R_i is different from their natural schedule of rights, it must be that their rights have changed in some way since birth. Three sorts of changes are possible: additions, subtractions, and replacements. Let $R'_i, R"_i$ be two logically possible schedules of rights. $R"_i$ can be obtained by addition from R'_i iff every $R'_{ij} \in R'_i$ is also an element of $R"_i$ and moreover there is exactly one $R"_{ij} \in R"_i$ that is not an element of R'_i. R'_i can be obtained by subtraction from $R"_i$ iff $R"_i$ can be obtained by addition from R'_i. At last, $R"_i$ is obtained from R'_i by replacement iff $|R'_i| = |R"_i|$ and there is one and only one pair of $R'_{ij} \in R'_i, R"_{ij} \in R"_i$ such that $R'_{ij} \neq R"_{ij}$. Now, define for each of $R'_i, R"_i$ a function $u'_i, u"_i$ in just the same way as before; I assume that the importance of a bundle of rights (e.g. the right to free speech) does not change if you gain or lose other rights, therefore I require that $u'_i(R'_{ij}) = u"_i(R"_{ij})$ iff $R'_{ij} = R"_{ij}$. Thus if some R'_i can be obtained by replacement, addition, or subtraction from $\overline{R}_i$, then $\overline{u}_i$ sets the origin and unit of u'_i. And since any individual i's actual assignment of rights R_i can be obtained from their natural schedule of rights $\overline{R}_i$ by a finite sequence of addition, subtraction, and replacement, $\overline{u}_i$ determines a unique function u_i representing the value of i's rights.

That was a long walk in the dark, but we are now in a position to propose a general measure of how deeply any given individual's rights have been disrespected overall. All else being equal, I believe it's twice as bad to infringe two of your rights than just one, and twice as bad to infringe one right twice as deeply, and twice as bad to infringe one right that is twice as important. Note that these statements are all meaningful: we can obviously always compare how many of your rights are violated with how many of mine are violated, and since all $d^x_{ij} \in [0,1]$ and all u_i share a common origin and unit, we can always make interpersonal comparisons of the depth of rights infringements and of the importance of the rights infringed. We *can* say "your right to free speech is twice as important as my right to bear arms, but is infringed only as third as deeply." Thus, let F be a family of $f_i : \{(I, A_x) | i \in I\} \longrightarrow \mathbb{R}$, representing the overall degree to which i's rights have been disrespected, and set $f_i(x) = \sum_j d^x_{ij} u_i(R_{ij})$. The construction is now complete. One can still haggle over details – for example, we could make the f_i concave over d^x_{ij} rather than linear – but in the main this process of construction gets us the family of functions we wanted.

Appendix F

Proof of Theorem 6.3.1

Proof. Necessity is obvious, so I only prove sufficiency. The proof model I adopt here closely follows that presented by Adler and McCarthy, adapted here only slightly for the framework I am using and for the variable-population context.

Assume there is a function $g : X \longrightarrow \mathbb{R}$ which satisfies A, SI, RI, PD, PI, C, and S. We start my making a conventional choice of common unit for all the f_i. By SI, no matter what choice we make, individual i's rights are overall disrespected to a greater degree in x than h's rights are disrespected in y iff $f_i(x) \geq f_h(y)$. Further, the difference in the degree to which i's rights are disrespected in x and individual h's rights are in y is at least as large as the difference between the degree to which individual k's rights are disrespected in z and individual l's rights are disrespected in zz iff $f_i(x) - f_h(y) \geq f_k(x) - f_l(y)$. Finally, the ratio of the degree to which i's rights are infringed to the degree to which h's rights are infringed is equal to a iff $\frac{f_i(x)}{f_h(y)} = a$.

Because $g(.)$ is separable and continuous, and orders X by the reals, established results in expected utility theory imply the existence of $\gamma_1, \gamma_2, ...$ such that

$$g(x) \geq g(y) \iff \sum_{i=1}^{n} \gamma_i(f_i(x)) \geq \sum_{i=1}^{N} \gamma_i(f_i(x)),$$

for n, N being the number of columns in $\mathbf{x}, \mathbf{y}$.

Since $g(.)$ is anonymous, $\gamma_i = \gamma_{i'}$ whenever $n = N$. Moreover since $g(.)$ satisfies RI, we know that if x replicates y then $\sum_{i=1}^{N} \gamma(f_i(x)) = \sum_{i=1}^{N} \gamma'(f_i(x))$ where N, n are the populations of x, y. There is a single γ for all i associated with the columns in x and a single γ' associated with the columns in y; this follows from my remark about anonymity. Note further that because x replicates y we know that $\sum_{i=1}^{N} f_i(x) = \frac{N}{n} \sum_{i=1}^{N} f_i(x)$. Therefore $\gamma(f_i(x))$ must be related to $\gamma'(f_i(y))$

by a factor of $\frac{n}{N}$. Simplifying slightly, there must therefore be a unique γ such that

$$g(x) \geq g(y) \iff \frac{1}{n}\sum_{i=1}^{n}\gamma(f_i(x)) \geq \frac{1}{N}\sum_{i=1}^{N}\gamma(f_i(x)),$$

for n, N being the populations of x, y.

Additionally, because $g(.)$ satisfies PI, γ must be increasing, and because γ satisfies PD, γ must be strictly concave. q.e.d. □

Appendix G

Proof of Theorem 6.3.2

The proof of theorem 4.2. runs in much the same way.

Proof. The proof of theorem 4.2. runs in much the same way. Necessity is obvious, so assume there is a function $g : X \longrightarrow \mathbb{R}$ which satisfies A,SI,RI,PD,PI,C, but not S. SI imposes the same constraint on $g(.)$ as before. Because $g(.)$ is non-separable and continuous and satisfies RI, results by J. Weymark (2018) and Fleurbaey (2018) establish that for any two $\mathbf{x}, \mathbf{y}$ there are $\beta(i)_{i \in n}$ $\alpha(i)_{i \in N}$ such that

$$g(x) \geq g(y) \iff \frac{1}{n} \sum_{i=1}^{n} \beta_i \cdot f_i(x) \geq \frac{1}{N} \sum_{i=1}^{N} \alpha \cdot f_i(x),$$

where n, N are the number of columns in $\mathbf{x}, \mathbf{y}$. Since $g(.)$ satisfies PD and PI, all the β_i and α_i are positive and can be arranged in a strictly increasing order such that a person whose rights are strictly more disrespected gets a greater factor β_i or α_i. Finally, since $g(.)$ is anonymous, $\{\beta_i\} = \{\alpha_i\}$ whenever $n = N$. q.e.d.

□

Bibliography

Adler, Matthew (2011). *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis*. Oxford University Press.

— (2018). "Prioritarianism: Room for Desert?" In: *Utilitas* 30 (2), pp. 172–197.

Ainslie, George (2001). *Breadkdown of Will*. Cambridge University Press.

Akerlof, George and Rachel Kranton (2000a). "Economics and Identity". In: *The Quarterly Journal of Economics* 115 (3), pp. 715–753.

— (2000b). "Economics and Identity". In: *The Quarterly Journal of Economics* 115.3, pp. 715–753.

Alexander, Larry (2000). "Deontology at the Threshold". In: *San Diego Law Review* 37, pp. 893–912.

Alkire, Sabina (2005). *Valuing Freedoms: Sen's Capability Approach and Poverty Reduction*. Oxford University Press.

Alkire, Sabina and James Foster (2007). "Counting and multidimensional poverty measurement". In: *OPHI Working Paper* 7.

Andreou, Chrishoula (2006). "Environmental Damage and the Puzzle of the Self-Torturer". In: *Philosophy & Public Affairs* 34 (1), pp. 95–108.

— (2015). "The real puzzle of the self-torturer: uncovering a new dimension of instrumental rationality". In: *Canadian Journal of Philosophy* 45 (5-6), pp. 562–575.

Anscombe, Elizabeth (1957). *Intention*. Oxford: Blackwell.

Arneson, Richard (1980). "Mill Versus Paternalism". In: *Ethics* 90, pp. 470–489.

— (1985). "Freedom and Desire". In: *Canadian Journal of Philosophy* 15.3, pp. 425–448.

Arrow, Kenneth (1963). *Social Choice and Individual Value*. New York: Wiley.

— (1995). "A Note on Freedom and Flexibility". In: *Choice, Welfare and Development: A Festschrift in Honour o/Amartya K. Sen*. Ed. by Kaushik Basu, Prasanta K. Pattanaik, and Kotaro Suzumura. Clarendon Press, pp. 7–16.

Atkinson, AB (1970). "On the Measurement of Inequality". In: *Journal of Economic Theory* 2 (3), pp. 244–263.

Bader, Ralf (2015). "Kantian Axiology and the Dualism of Practical Reason". In: *Oxford Handbook of Value Theory*. Ed. by Iwao Hirose and Jonas Olson. Oxford University Press.

Bader, Ralf (2018). "Moralizing Liberty". In: *Oxford Studies in Political Philosophy* 4, pp. 141–166.

— (2019). "Agent-Relative Prerogatives and Suboptimal Beneficence". In: *Oxford Studies in Normative Ethics, Volume 11*. Ed. by Mark Timmons. Oxford University Press.

Barberà, Salvador, Walter Bossert, and Prasanta K. Pattanaik (2005). "On Ranking Sets of Objects". In: *Handbook of Utility Theory*. Ed. by Salvador Barberà, J. P. Hammond, and C. Seidl. Berlin: Springer.

Barberá, Salvador and Brigit Godal (2010). "Preference for flexibility and the opportunities of choice". In: *Journal of methematical economics* 47, pp. 272–278.

Baumeister, R. F. and T. Heatherton. (1996). "Self-regulation failure: An overview". In: *Psychological Inquiry* 7, pp. 1–15.

Baumgärtner, Stefan (2006). *Measuring the Diversity of What? And for What Purpose? A Conceptual Comparison of Ecological and Economic Biodiversity Indices*. Mimeo. Department of Economics, University of Heidelberg.

Bedau, Hugo Adam (2004). "Capital Punishment". In: *The Oxford Handbook of Practical Ethics*. Ed. by Hugh LaFollette. Oxford University Press, pp. 705–733.

Bem, D. J. (1972). "Self-perception theory". In: *Advances in experimental social psychology*. Ed. by L. Berkowitz. Vol. 6. New York: Academic Press.

Bentehm, Johan van and Fenrong Liu (2007). "Dynamic Logic of Preference Upgrade". In: *Journal of Applied Non-Classical Logics* 17 (2), pp. 157–182.

Berlin, Isiah (1969). "Four Essays on Liberty". In: Oxford: Oxford University Press. Chap. Two Conceptions of Liberty, pp. 118–172.

Bervoets, Sebastien and Nicolas Gravel (2007). "Appraising diversity with an ordinal notion of similarity: An axiomatic approach". In: *Mathematical Social Sciences* 53, pp. 259–273.

Binmore, Ken (2008). *Rational Decisions*. Princeton, NJ: Princeton University Press.

Bodner, Ronit and Drazen Prelec (2002). "Self-signaling and diagnostic utility in everyday decision making". In: *Collected Essays in Psychology and Economics*. Ed. by I. Brocas and J. Carillo. Oxford: Oxford University Press.

Bossert, William, Prasanta K. Pattanaik, and Yongsheng Xu (1992). "Similarity of Options and the Measurement of Diversity". In: *Journal of Theoretical Politics* 15, pp. 405–21.

Bovens, Luc (2009). “The Ethics of Nudge”. In: *Preference Change: Approaches From Philosophy, Economics and Psychology.* Ed. by Till Grüne-Yanoff and Sven Ove Hansson. Berlin: Springer, pp. 207–219.

— (2013). “Essays in Behavioural Public Policy”. In: *Essays in Behavioural Public Policy.* Ed. by Adam Oliver. Cambridge: Cambridge University Press, pp. 228–233.

Bowen, S. et al. (2006). “Mindfulness meditation and substance use in an incarcerated population”. In: *Psychology of Addictive Behaviors* 20.3, pp. 343–347.

Bowie, Lee G. (1979). “The Similarity Approach to Counterfactuals: Some Problems”. In: *Noûs* 13 (4), pp. 477–498.

Bradley, Richard (2017). *Decision Theory with a Human Face.* Cambridge University Press.

— (2018). “Decision Theory: A Formal Philosophical Introduction”. In: *Introduction to Formal Philosophy.* Ed. by S Hansson and V Hendricks. Springer.

Bradley, Richard and Christian List (2009). “Desire-as-Belief Revisited”. In: *Analysis* 69 (1), pp. 31–37.

Bradley, Richard and Orri Stefánsson (2010). “Desire, Belief, and Invariance”. In: *Mind* 125 (499), pp. 691–725.

Brams, Stephen J and Peter C Fishburn (1978). “Approval Voting”. In: *The American Political Science Review* 72 (3), pp. 831–847.

Briggs, Rachael (2015). “Costs of Abandoning the Sure-Thing Principle”. In: *Canadian Journal of Philosophy* 45 (5-6), pp. 827–840.

Broad, C D (1933). “Is ‘Goodness’ a Name of a Simple Non-Natural Quality?” In: *Proceedings of the Aristotelian Society, New Series* 34, pp. 249–68.

Broome, John (1991a). “Desire, Beliefs, and Expectation”. In: *Mind* 100 (2), pp. 265–267.

— (1991b). “Utility”. In: *Exconomics and Philosophy* 7, pp. 1–12.

Brooome, John (1999a). “Can a Humean be Moderate”. In: *Ethics out of Economics.* Oxford: Oxford University Press, pp. 68–87.

— (1999b). *Weighing Lives.* Oxford: Oxford University Press.

Brown, Campbell (2011). “Consequentialize This”. In: *Ethics* 121 (4), pp. 749–771.

Buchak, Lara (2013a). “Free Acts and Change: Why the Rollback ARgument Fails”. In: *The Philosophical Quarterly* 63 (250), pp. 20–28.

— (2013b). *Risk and Rationality.* Oxford University Press.

Callard, Agnes (2018). *Aspiration: the Agency of Becoming.* London: Oxford University Press.

Carlson, Erik (2006). “Cyclical Preferences and Rational Choice”. In: *Theoria* 62 (1-2), pp. 144–160.

Carter, Ian (1999). *A Measure of Freedom*. Oxford: Oxford University Press.

— (2004). “Choice, Freedom, and Freedom of Choice”. In: *Social Choice and Welfare* 22, pp. 61–81.

Carter, Ian and Ronen Shnayderman (2019). “The Impossibility of “Freedom as Independence””. In: *Political Studies Review* 17 (2), pp. 136–146.

Chiesa, A and A Serretti (2014). “Are Mindfulness-Based Interventions Effective for Substance Use Disorders? A Systematic Review of the Evidence”. In: *Substance Use & Misuse* 49.5.

Christman, John (2001). “Liberalism and Individual Positive Freedom”. In: *Ethics* 101, pp. 343–359.

Christopher, Russel L (2003). “The Proscutor’s Dilemma: Bargains and Punishments”. In: *Fordham Law Review* 72 (1), pp. 893–912.

Cohen, G A (1978). “Capitalism, Freedom, and the Proletariat”. In: *The Idea of Freedom: Essays in Honor of Isiah Berlin*. Ed. by A. Ryan. London: Oxford University Press.

Cohen, G. A. (2011). “Capitalism, Freedom, and the Proletariat”. In: *On the Currency of Egalitarian Justice, and Other Essays in Political Philosophy*. New York: Routledge, pp. 147–165.

Crocker, Lawrence (1980). *Positive Liberty*. London: Nijhoff.

Davidson, Donald (1969). “How is Weakness of the Will Possible?” In: *Moral Concepts*. Ed. by Joel Feinberg. Oxford University Press.

— (1980). *Essays on Actions and Events*. Oxford: Clarendon Press.

Debreu, Gérard (1954). “Representation of a Preference Ordering by a Numerical Function”. In: *Decision Processes*. Ed. by Robert M Trhall, Clyde H Coombs, and Howard Raiffa. Wiley, pp. 159–167.

Dostoyevsky, Fyodor (1864). *Notes from the Underground*. Electronic copy from Planet eBook.com. URL: https://www.planetebook.com/free-ebooks/notes-from-the-underground.pdf.

Dowding, Keith and Martin Van Hees (2009). “Freedom of Choice”. In: *The Handbook of Rational and Social Choice*. Ed. by Paul Anand, Prasanta K. Pattanaik, and Clemens Puppe. Oxford: Oxford University Press, pp. 374–390.

Duff, Anthony (2001). *Punishment, communication, and community*. Oxford University Press.

Duus-Otterström, Göran and Erin l. Kelly (2019). “Injustice and the Right to Punish”. In: *Philosophy Compass* 11 (3).

Dworkin, Gerald (1982). “Is More Choice Better than Less?” In: *Midwest Studies In Philosophy* 7, pp. 47–61.

Dworkin, Ronald (1977). *Taking Rights Seriously*. Harvard University Press.

— (2000). *Sovereign Virtue.* Harvard University Press.

— (2011). *Justice for Hedgehogs.* Harvard University Press.

Edgington, Dorothy (1978). "What If? Questions About Conditionals?" In: *Mind & Language* 18 (4), pp. 380–401.

Elkins-Brown, N, R Teper, and M Inzlicht (2017). "How Mindfulness Enhances Self-Control". In: *Mindfulness in Social Psychology.* Ed. by J.C. Karremans and E.K. Papies. New York: Psychology Press.

Ellis, Anthony (1992). "Deontology, Incommensurability and the Arbitrary". In: *Philosophy and Phenomenological Research* 52 (4), pp. 855–875.

Ellsberg, Daniel (1954). "Classic and Current Notions of 'Measurable Utility'". In: *The Economic Journal* 64.255, pp. 528–556.

Elwafi, H. M. et al. (2013). "Mindfulness training for smoking cessation: Moderation of the relationship between craving and cigarette use". In: *Drug and Alcohol Dependence* 130.1-3, pp. 222–229.

Flathman, Richard (1987). *The Philosophy and Politics of Freedom.* Chicago University Press.

Fleurbaey, Marc (2018). "Assessing Risky Social Situations". In: *Journal of Political Economy* 118, pp. 649–80.

Foster, James (2011). "Freedom, Opportunity, and Well-being". In: *Handbook of Social Choice and Welfare.* Vol. 2. Elsevier BV, pp. 687–728.

Foweraker, Joe and Todd Landman (2000). *Citizenship Rights and Social Movements: A Comparative and Statistical Analysis.* Oxford University Press.

Friedman, Milton (1962). *Capitalism and Freedom.* Chicago: University of Chicago Press.

Fudenberg, Drew and David K. Levine (2012). "Timing and Self-Control". In: *Econometrica* 80, pp. 1–42.

Gaertner, Wulf, Prasanta K. Pattanaik, and Kotaro Suzumura (1992). "Individual Rights Revisited". In: *Economica* 59, pp. 161–77.

Gärdenfors, Peter and Nils-Eric Sahlin (1982). "Unreliable probabilities, risk taking, and decision makin". In: *Synthese* 53 (3), pp. 361–386.

Garnett, Michael (2007). "Ignorance, Incompetence, and the Concept of Liberty". In: *Journal of Political Philosophy* 15 (4), pp. 361–386.

— (2016). "VALUE NEUTRALITY AND THE RANKING OF OPPORTUNITY SETS". In: *Economics and Philosophy* 16, pp. 99–119.

Gastil, Raymond Duncan (1990). "The Comparative Survey of Freedom: Experiences and Suggestions". In: *Studies in Comparative International Development* 25, pp. 25–50.

Gibbard, Alan (1996). “A Pareto-Consistent Libertarian Claim”. In: *Journal of Economic Theory* 7 (4), pp. 288–410.

Gilboa, Itzhak and David Schmeidler (1982). “Maxmin expected utility with non-unique prior”. In: *Journal of Mathematical Economics* 18 (2), pp. 141–153.

Goldman, Alvin (1970). *A Theory of Human Action.* Englewood Falls: Prentice Hall.

Goodman, Jeremy (2015). *Counterfactuals and comparative similarity.* Working draft: http://www-bcf.usc.edu/ jlgoodma/CounterfactualsWithoutCloseness.pdf.

Greaves, Hilary (2017). “A Reconsideration of the Harsanyi–Sen–Weymark Debate on Utilitarianism”. In: *Utilitas* 29 (4), pp. 175–213.

Gustafsson, Johan E. (2019). “A Paradox for the Intrinsic Value of Freedom of Choice”. In: *Noûs* ISSN 1468-0068, pp. 1–23.

Hagger, Martin, L Nikos, and D Chatzisarantis (2016). “A Multilab Preregistered Replication of the Ego Depletion Effect”. In: *Perspectives on Psychological Science* 11, pp. 546–573.

Hammond, Peter J (1976). “Changing Tastes and Coherent Dynamic Choice”. In: *The Review of Economic Studies* 43, pp. 159–173.

Hansson, Sven Ove (1995). “Changes in Preference”. In: *Theory and Decision* 38 (2), pp. 1–28.

Harsanyi, John (1975). “Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls’s Theory”. In: *American Political Science Review* 69, pp. 594–606.

Hausman, Dan and Brynn Welch (2010). “Debate: To Nudge or Not to Nudge”. In: *The Journal of Political Philosophy* 18.1, pp. 123–136.

Hay, Carol (2013). *Kantianism, Liberalism, and Femism: Resisting Oppression.* Palgrave Macmillan.

Hayek, Friedrich (1944). *The Road to Serfdom.* Chicago: Routledge.

— (1960). *The Constitution of Liberty.* Chicago: University of Chicago Press.

Hees, Martin Van (2004). “Freedom of choice and diversity of options: Some difficulties”. In: *Social Choice and Welfare* 22, pp. 253–266.

Hicks, J R (1939). *Value and capital: An inquiry into some fundamental principles of economic theory.* Oxford: Clarendon Press.

Hill, Brian (2013). “Confidence in Preferences”. In: *Social Choice and Welfare* 39, pp. 273–302.

Hirose, Iwao (2014). *Moral Aggregation.* Oxford University Press.

Hobbes, Thomas (1994). *Leviathan.* Hackett.

Holton, Richard (2006). *Willing, Wanting, Waiting.* Oxford: Oxford University Press.

Holton, Richard and Kent C. Berridge. (2013). “Compulsion and Choice in Addiction”. In: *Addiction and Self-Control.* Ed. by Neil Levy. Oxford: Oxford University Press.

Hurka, Thomas (2011). *Drawing Morals: Essays in Ethical Theory.* Oxford: Oxford University Press.

Inwagen, Peter van (2000). “Free Will Remains a Mystery”. In: *Philosophical Perspectives* 13, pp. 1–20.

Inzlicht, Michael and Brandon Schmeichel (2012). “What is Ego Depletion? Towards a Mechanistic Revision of the Resource Model of Self-Contro”. In: *Perspectives on Psychological Science* 7, pp. 450–463.

James, William (1980). *The Principles of Psychology.* Vol. 2. New York: Henry Holt and Company.

Jeffrey, Richard (1965). *The Logic of Decision.* Chicago: University of Chicago Press.

Johnson, Christa (2018). “The Intrapersonal Paradox of Deontology”. In: *Journal of Moral Philosophy*, pp. 279–301.

— (2020). “How Deontologists Can Be Moderate (and Why They Should Be)”. In: *Journal of Value Inquiry.* Pre-published.

Kahneman, Daniel and Amos Tversky (1979). “Prospect Theory: An Analysis of Decision Under Risk”. In: *Econometrica* 47 (2), pp. 263–292.

Kamm, Frances (1995). “Inviolability”. In: *Midwestern Studies in Philosophy*, pp. 165–175.

— (1996). *Morality, Mortality, vol II.* Oxford University Press.

Karni, Edi and Marie-Louise Vierø (2013). “Reverse Bayesianism: A choice-based theory of growing awareness”. In: *American Economic Review* 103 (7), pp. 2790–2810.

Keynes, John Maynard (1973/1921). *A Treatise on Probability, vol. VIII, The Collected Writings of John Maynard Keynes.* London: Macmillan.

Khader, Serene (2011). *Adaptive Preferences and Women’s Empowerment.* Oxford: Oxford University Press.

Kim, S. D. (2012). *Characterising unknown unknowns.* Paper presented at PMI Global Congress 2012—North America, Vancouver, British Columbia, Canada. Newtown Square, PA: Project Management Institute.

Klemish-Ahlert, M (1993). “Freedom of choice. A comparison of different rankings of opportunity Sets”. In: *Social Choice and Welfare* 10, pp. 289–173.

Klemish-Ahlhert, M. (2004). “Freedom of choice. A comparison of different rankings of opportunity Sets”. In: *Social Choice Welfare* 10, pp. 189–173.

Klocksiem, Justin (2015). "How to Accept the Transitivity of Better Than". In: *Philosophical Studies* 173, pp. 1309–1334.

Kment, Boris (2006). "Counterfactuals and Explanation". In: *Mind* 115 (458), pp. 261–310.

Kramer, Matthew (2003). *The Quality of Freedom.* Oxford: Oxford University Press.

Kreps, D. M. (1979). "A Representation Theorem for "Preference for Flexibility"". In: *Econometrica* 47, pp. 565–77.

Kristjánsson, Kristjan (1996). *Social Freedom: The Responsibility View.* Cambridge: Cambridge University Press.

Krodel, Thomas and Franz Huber (2013). "Counterfactual Dependence and Arrow". In: *Noûs* 47 (3), pp. 453–466.

Kukathas, Chandran (2003). *The Liberal Archipelago: A Theory of Diversity and Freedom.* Oxford: Oxford University Press.

Kymlicka, Will (1990). *Contemporary Political Philosophy.* Oxford: Oxford University Press.

Landman, Todd (2004). "Measuring Human Rights". In: *Human Rights Quarterly* 26 (4), pp. 906–931.

Lenman, James (2000). "Consequentialism and Cluelessness". In: *Philosophy & Public Affairs* 29 (4), pp. 342–370.

Levi, Isaac (1990). *Hard Choices: Decision Making under Unresolved Conflict.* Cambridge University Press.

Lewis, David (1979). "Counterfactual Dependence and Time's Arrow". In: *Noûs* 13 (4), pp. 455–476.

List, Christian and Laura Valentini (2016). "Freedom as Independence". In: *Ethics* 126, pp. 1043–1074.

Loch, C. H., M. E. Solt, and E. M. Bailey (2007). "Diagnosing unforeseeable uncertainty in a new venture". In: *Journal of Product Innovation Management* 25, pp. 28–56.

Locke, John (1948). *The Second Treatise of Civil Government and A Letter Concerning Toleration.* Blackwell.

MacCallum, G C (1967). "Negative and Positive Freedom". In: *Philosophical Review* 76, pp. 312–24.

Manley, David and Ryan Wasserman (2008). "How is Weakness of the Will Possible?" In: *Mind* 111, pp. 59–84.

Mann, Traci and Andrew Ward (2007). "Attention, Self-Control, and Behaviour". In: *Current Directions in Psychological Science* 16, pp. 280–283.

Marks, Louis, Tony Smith, and Oscar Wilde (1987). "Lady Windermere's Fan". In: *BBC.*

McCarthy, David (2018). "Risk-Free Approaches to the Priority View". In: *Erkenntnis* 78, pp. 421–49.

McClennen, James (1990). *Rationality and Dynamic Choice.* Cambridge University Press.

McElwee, Brian (2017). "Demandingness Objections in Ethics". In: *The Philosophical Quarterly* 67 (266), pp. 84–105.

Merrill, Samuel (1979). "Approval Voting: A 'Best Buy' Method for Multicandidate Elections?" In: *Mathematics Magazine* 52 (2), pp. 98–102.

Mill, John Stuart (1885). *On Liberty.* Batoche Books.

Miller, D (1983). "Constraints on Freedom". In: *Ethics* 94, pp. 66–86.

Moore, Michael S. (1989). "Torture and the Balance of Evils". In: *Israel Law Review* 23, pp. 280–344.

Morreau, Matthew (2004). "It Simply Does Not Add Up: Trouble With Overall Similarity". In: *Journal of Philosophy* 107 (9), pp. 469–490.

Muller, Edward N. and Mitchell A. Seligson (1987). "Inequality and Insurgency". In: *The American Political Science Review* 81 (2), pp. 425–51.

Munoz-Dardé, Véronique (2015). "The Quality of Gooditude". In: *Journal of Moral Philosophy* 12 (4), pp. 393–413.

Muraven, M, M. T. Diane, and Roy F. Baumeister (1998). "Self-Control as Limited Resource: Regulatory Depletion Patterns". In: *Journal of Personality and Social Psychology* 74, pp. 774–789.

Nagel, Thomas (1972). "War and Massacre". In: *Philosophy & Public Affairs* 1 (2), pp. 123–144.

— (1975). "Libertarianism without Foundations". In: *The Yale Law Journal* 85 (1), pp. 136–149.

— (1980). "The Limits of Objectivity". In: *The Tanner Lectures on Human Values, vol. I.* Ed. by S. McMurrin. Cambridge University Press.

— (2007). *Morality and Self-Interest.* Oxford University Press.

Al-Najjar, Nabil I. and Jonathan Weinstein (2009). "The Ambiguity Aversion Literature: A Critical Assessment". In: *Economics and Phiosophy* 23 (3).

Nebel, Jacob (2018). "The Good, the Bad, and the Transitivity of Better Than". In: *Noûs* 52 (4), pp. 874–899.

Nehring, Klaus and Clemens Puppe (2009). "Diversity". In: *The Handbook of Rational and Social Choice.* Ed. by Paul Anand, Prasanta K. Pattanaik, and Clemens Puppe. Oxford: Oxford University Press, pp. 298–318.

Nickel, James (2001). “Ian Carter, A Measure of Freedom”. In: *Law and Philosophy* 20, pp. 531–540.

Nietzsche, Friedrich (1968). “Twilight of the Idols”. In: ed. by R. J. Hollingdale. Baltimore: Penguin.

Norman, R. (1978). “Liberty, Equality, Property”. In: *Proceedings of the Aristotelian Society Suppl. Vol.* 55, pp. 193–209.

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

Nussbaum, Martha (2001). “Adaptive Preferences and Women’s Options”. In: *Economics and Philosophy* 17, pp. 67–88.

O’Neill, Onora (1980). “Matters of Life and Death”. In: *Kantian Approaches to Some Famine Problems.* Ed. by T. Regan. McGraw-Hill Companies, pp. 546–551.

Oppenheim, Felix (1981). *Political Concepts: A Reconstruction.* Oxford: Blackwell.

— (1995). “Social Freedom and its Parameters”. In: *Journal of Theoretical Politics* 7.4, pp. 403–420.

— (2004). “Social Freedom: Definition, Measurement, and Valuation”. In: *Social Choice and Welfare* 22.4, pp. 175–185.

Otsuka, Michael (2003). *Libertarianism Without Inequality.* Oxford University Press.

Parfit, Derek (1981). *On What Matters, V 1.* Oxford: Oxford University Press.

Pattanaik, Prasanta K. (2008). “Rights, Individual Preferences, and Collective Rationality”. In: *Arguments for a Better World: Essays in Honor of Amartya Sen: Volume I: Ethics, Welfare, and Measurement.* Ed. by Kaushik Basu and Ravi Kanbur. Oxford: Oxford University Press, pp. 68–79.

Pattanaik, Prasanta K. and Kotaro Suzumura (1996). “Individual Rights and Social Evaluation: A Conceptual Framework”. In: *Oxford Economic Papers* 48 (2), pp. 194–212.

Pattanaik, Prasanta K. and Yongsheng Xu (1990). “On Ranking Opportunity Sets in Terms of Freedom of Choice”. In: *Recherches Économiques de Louvain* 56, pp. 383–390.

— (1998). “On Preference and Freedom”. In: *Theory and Decision* 44, pp. 173–98.

— (2000). “On diversity and freedom of choice”. In: *Mathematical Social Sciences* 20, pp. 123–130.

— (2002). “Minimal relativism, dominance, and standard of living comparisons based on functionings”. In: *Oxford Economic Papers* 59, pp. 354–374.

— (2008). “Ordinal Distance, Dominance, and the Measurement of Diversity”. In: *Rational Choice and Social Welfare.* Ed. by Prasanta K. Pattanaik et al. Berlin: Springer, pp. 259–269.

— (2013). *Rationality and Context-Dependent Preferences.* Andrew Young School of Policy Studies Research Paper Series No. 13-13. URL: https://ssrn.com/abstract=2327260.

— (2015). "Freedom and its Value". In: *The Oxford Handbook of Value Theory.* Ed. by I. Hirose and J. Olson. Oxford: Oxford University Press.

Paul, L.A. (2014). *Transformative Experience.* London: Oxford University Press.

Pearl, Judea (2000). "Causality: Models, Reasoning and Inference". In: *Tijdschrift Voor Filosofie* 64 (1), pp. 201–202.

Peterson, Martin and Peter Vallentyne (2018). "Self-Prediction and Self-Control". In: *Self-Control, Decision Theory, and Rationality.* Ed. by Jose Luis Bermudez. Cambridge University Press, pp. 48–71.

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

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

— (2003). "Akrasia, Collective and Individual". In: *Weakness of Will and Practical Irrationality.* Ed. by S. Stroud and C. Tappolet. Oxford University Press.

— (2012). *Just Freedom: A Moral Compass for Society.* London and New York: WW Norton & Company.

— (2014). *On the People's Terms: A Republican Theory and Model of Democracy.* Cambridge: Cambridge University Press.

Pettit, Philip and Kinch Hoekstra (2018). *The Birth of Ethics.* Oxford: Oxford University Press.

Ponthière, Gregory (2003). "Utilitarian Population Ethics: a Survey". In: *CREPP Working Papers*, p. 0303.

Popper, Karl (1959/1934). *The Logic of Scientific Discovery.* London: Hutchinson.

Portmore, Douglas W. (2001). "Can an Act-Consequentialist Theory be Agent Relative?" In: *American Philosophical Quarterly* 38 (4), pp. 363–377.

Puppe, Clemens (1996). "An Axiomatic Approach to 'Preference for Freedom of Choice'". In: *Journal of Economic Theory* 68, pp. 174–199.

Quinn, Warren (1994). "The right to threaten and the right to punish". In: *Morality and Action.* Cambridge Studies in Philosophy. Cambridge University Press, pp. 52–100. DOI: 10.1017/CBO9781139172677.004.

Rabinowicz, Wlodek (1995). "To Have One's Cake and Eat it Too: Sequential Choice and Expected-utility Violations". In: *Journal of Philosophy* 92 (11), pp. 586–620.

Rachels, Stuart (1998). "Counterexamples to the Transitivity of Better Than". In: *Australasian Journal of Philosophy* 76 (1), pp. 71–83.

Rawls, John (1971). *A Theory of Justice*. Harvard University Press.

Raz, Joseph (1988). *The Morality of Freedom*. Oxford: Oxford University Press.

Reporters Without Borders (2019). *2019 World Press Freedom Index*. Annual Report. URL: https://rsf.org/en/ranking_table.

Rizzo, Mario J (2016). "Behavioral Economics and Deficient Willpower: Searching for Akrasia". In: *SSRN Electronic Journal, https://ssrn.com/abstract=2731818*.

Robinson, Terry E and Kent C. Berridge (2003). "Addiction". In: *Annual Review of Psychology* 54, pp. 25–53.

Rosenbaum, E. F. (2000). "On Measuring Freedom". In: *Journal of Theoretical Politics* 12, pp. 205–77.

Salib, Peter. "Why Prison?: An Economic Critique". In: *Berkeley Journal of Criminal Law 2017*. URL: https://ssrn.com/abstract=2928219. Forthcoming.

Samuelson, Paul (1937). "A Note on Measurement of Utility". In: *Review of Economic Studies* 4 (2), pp. 451–485.

— (1938). "A note on the pure theory of consumer's behaviour". In: *Econometrica* 5 (17), pp. 61–71.

Savage, Leonard J (1954). *The Foundations of Statistics*. John Wiley and Sons.

Scanlon, T. M. (1998). *What we Owe to Each Other*. Harvard University Press.

— (2008). "Rights and Interests". In: *Arguments for a Better World: Essays in Honor of Amartya Sen: Volume I: Ethics, Welfare, and Measurement*. Ed. by Kaushik Basu and Ravi Kanbur. Oxford: Oxford University Press.

Scheffler, Samuel (1985). "Agent-Centred Restrictions, Rationality, and the Virtues". In: *Mind* 94, pp. 409–419.

Sen, Amartya (1970). "The Impossibility of a Paretian Liberal". In: *Journal of Political Economy* 78, pp. 152–157.

— (1971). "Choice Functions and Revealed Preference". In: *The Review of Economic Studies* 38 (3), pp. 307–317.

— (1976). "Welfare inequalities and Rawlsian axiomatics". In: *Theory and decision* 7, pp. 243–262.

— (1977). "On Weights and Measures: Informational Constraints in Social Welfare Analysis". In: *Econometrica* 45, pp. 1539–1572.

— (1982). "Rights and Agency". In: *Philosophy and Public Affairs* 11 (3), pp. 3–39.

— (1986). "Social Choice Theory". In: *Handbook of Mathematical Economics: volume III*. Ed. by Kenneth Arrow and Michael Intriligator. North Holland, pp. 1073–1181.

— (1990a). “Development as Capability Expansion”. In: *Human Development and the International Development Strategy for the 1990s.* Ed. by K. Griffin and J. Knight. London: Macmillan.

— (1990b). “Gender and cooperative conflicts”. In: *Persistent Inequalities.* Ed. by I. Tinker. Oxford: Oxford University Press.

— (1990c). “Welfare, Freedom, and Social Choice: A Reply”. In: *Recherches Économiques de Louvain* 56, pp. 451–85.

— (1993a). “Internal Consistency of Choice”. In: *Econometrica* 61 (3), pp. 495–521.

— (1993b). “Markets and Freedoms: Achievements and Limitations of the Market Mechanism in Promoting Individual Freedoms”. In: *Oxford Economic Papers*, pp. 519–541.

— (1999). *Development as Freedom.* Knopf Press.

— (2002). *Rationality and Freedom.* Harvard University Press.

— (2009). *The Idea of Justice.* London: Penguin Books.

Sen, Amartya, Kenneth Arrow, and Kotaro Suzumura (2011). “Kenneth Arrow on Social Choice Theory”. In: *Handbook of Social Choice and Welfare.* Vol. 2. Elsevier BV.

Skyrms, Brian (1977). “Resiliency, Propensity, and Causal Necessity”. In: *The Journal of Philosophy* 74.11, pp. 704–713.

Slote, Michael (1978). “Time in Counterfactuals”. In: *Philosophical Review* 87 (1), pp. 3–27.

Smith, Michael (2003). “Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion”. In: *Weakness of Will and Practical Irrationality.* Ed. by S. Stroud and C. Tappolet. Oxford University Press.

Srinivasan, Amia (2013). “Are we Luminous?” In: *Philosophy and Phenomenological* 90 (2), pp. 294–319.

Stalnaker, R. (1968). “A Theory of Conditionals”. In: *American Philosophical Quarterly, monograph series* 2, pp. 98–112.

Stanley, Matthew L., Gregory W. Stewart, and Felipe De Brigard (2017). “Counterfactual Plausibility and Comparative Similarity”. In: *Cognitive Science* 41, pp. 1216–1228.

Steiner, Hillel (1977). “The Structure of a Set of Compossible Rights”. In: *Journal of Philosophy* 74 (12), pp. 767–775.

— (1983). “How Free: Computing Personal Liberty”. In: *Royal Institute of Philosophy Lectures*, pp. 73–89.

— (1994). *An Essay on Rights.* Oxford: Blackwell.

Stepan, Alfred and Cindy Skach (1993). “Constitutional Frameworks and Democratic Consolidation: Parliamentarianism versus Presidentialism”. In: *World Politics* 46 (1), pp. 1–22.

Stroud, Sarah (2003). “Weakness of Will and Practical Judgment”. In: *Weakness of Will and Practical Irrationality*. Ed. by S. Stroud and C. Tappolet. Oxford University Press.

Sudgen, Robert and Peter Jones (1982). “Evaluating Choice”. In: *International Review of Law and Economics* 2 (1), pp. 47–65.

Sugden, Robert (1985). “Liberty, Preference, and Choice”. In: *Economics and Philosophy* 1, pp. 213–229.

— (1992). “The Metric of Opportunity”. In: *Economics and Philosophy* 14, pp. 307–337.

— (1996). “Rational Choice: A Survey of Contributions from Economics and Philosophy”. In: *The Economic Journal* 101 (407), pp. 751–785.

Sunstein, Cass (2019). *On Freedom*. Princeton University Press.

Suppes, Patrick (1987). “Maximizing Freedom of Decision: An Axiomatic Analysis”. In: *Arrow and the Foundations of the Theory of Economic Policy*. Ed. by George R. Feiwel. London: Palgrame MacMillan, pp. 243–254.

— (1996). “The Nature and Measurement of Freedom”. In: *Social Choice and Welfare* 13, pp. 183–200.

Suzumura, Kotaro and Naoki Yoshihara (2006). *On Initial Conferment of Individual Rights*. Discussion Paper Series a478. Institute of Economic Research, Hitotsubashi University. URL: https://EconPapers.repec.org/RePEc:hit:hituec:a478.

Tang, Y Y, R Tang, and M I Posner (2013). “Brief meditation training induces smoking reduction”. In: *Proceedings of the National Academy Sciences* 110.34, pp. 13971–13975.

Taylor, Charles (1985). “What’s Wrong With Negative Liberty”. In: *Philosophy and the Human Sciences: Philosophical Papers*. Vol. 2. Cambridge: Cambridge University Press, pp. 211–29.

Temkin, Larry (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford University Press.

— (2015). “Rethinking Rethinking the Good”. In: *The Journal of Moral Philosophy* 12, pp. 479–538.

Tennenbaum, Sergio (2014). “The Perils of Earnest Consequentializing”. In: *Philosophy and Phenomenological Research* 88 (1), pp. 233–240.

— (2017). “Action, Deontology, and Risk: Against the Multiplicative Model”. In: *Ethics*, pp. 674–707.

Thaler, Richard H and Shlomo Benartzi (2003). “Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving”. In: *Journal of Political Economy* 112, S164–S187.

Thaler, Richard H and Cass Sunstein (2003). “Libertarian Paternalism”. In: *The American Economic Review* 90, pp. 175–179.

The Fraser Institute (2019). *The Human Freedom Index 2019*. Annual Report. URL: https://www.fraserinstitute.org/studies/human-freedom-index-2019.

Thoma, Johanna. “In Defense of Revealed Preference Theory”. In: *Economics and Philosophy*. URL: https://johannathoma.files.wordpress.com/2020/02/in-defence-of-revealed-preference-theory-final-eap.pdf?fbclid=IwAR1ybrejlXeb84U2W1tf0tjkgGwYkOo_csu-14Lbt6vMTea_N3y5ztIf2gk. Forthcoming.

Thomson, Judith Jarvis (1984). *The Realm of Rights*. Harvard University Press.

— (1992). *The Realm of Rights*. Harvard University Press.

Tversky, Amos (1969). “Intransitivity of preferences”. In: *Psychological Review* 76, pp. 31–48.

Vallentyne, Peter (2002). “Brute Luck, Option Luck, and Equality of Initial Opportunities”. In: *Ethics* 112, pp. 529–557.

Vallentyne, Peter, Hillel Steiner, and Michael Otsuka (2005). “Why Left-Libertarianism Is Not Incoherent, Indeterminate, or Irrelevant: A Reply to Fried”. In: *Philosophy & Public Affairs* 33 (2), pp. 201–215.

Van Hees, Martin (2000). *Legal Reductionism and Freedom*. Kluwer Academic Publishers.

Velleman, David (2015). “Against the Right to Die”. In: Open Book Publishers.

Vohs, Kathleen D and Roy Baumeister (2004). *Handbook of self-regulation: Research, theory, and applications*. The Gilford Press.

Vohs, Kathleen D and Todd F Heatherton (2000). “Self-Regulatory Failure: A Resource-Depletion Approach”. In: *Psychological Science* 11, pp. 249–259.

Wasserman, Ryan (2006). “The Future Similarity Objection”. In: *Synthese* 150 (1), pp. 57–67.

Weitzman, M. (1992). “On Diversity”. In: *Quarterly Journal of Economics* 207, pp. 363–405.

Weymark, John (2000). “Measure Theory and the Foundations of Utilitarianism”. In: *Social Choice and Welfare* 25, pp. 527–555.

— (2018). “Generalized Gini Inequality Indices”. In: *Mathematical Social Science* 1, pp. 409–30.

Weymark, John A. (1991). “Reconsideration of the Harsanyi-Sen Debate on Utilitarianism”. In: *Interpersonal Comparisons of Well-Being*. Ed. by Jon Elster and John Roemer. Cambridge University Press, pp. 255–320.

Wilde, Oscar (1891). *The Picture of Dorian Gray*. London: Lippincott’s Monthly Magazine.

Williams, Bernard (1974). “A Critique of Utilitarianism”. In: *Utilitarianism: For and Against*. Ed. by J.J.C Smart and Bernard Williams. Cambridge University Press.

Williamson, Timothy (1992). “Vagueness and Ignorance”. In: *Proceedings of the Aristotelian Society, Supplementary Volumes* 66, pp. 145–177.

Xu, Yongsheng and Clemens Puppe (2000). “Essential Alternatives and Freedom Rankings”. In: *Social Choice and Welfare* 35, pp. 669–685.

Zamir, Eyal and Barak Medina (2010). *Law, Economics, and Morality*. Oxford University Press.