Feasibility Constraints for Political Theories

Holly Lawford-Smith

November, 2010

A thesis submitted for the degree of Doctor of Philosophy of The Australian National University.

Declaration

With the exception of the sections listed below, the work in this thesis is original and completely my own. Sections 4.3.1, 4.3.3, 5.3, 5.4.1, 5.4.3 and 6.3 draw upon parts of a co-authored paper with Pablo Gilabert, for which our contributions to the research were equal. No part of this thesis has been submitted for any degree or is currently being submitted for any other degree. To the best of my knowledge, all help received in preparing this dissertation, and all sources used, have been acknowledged.

Signed _____ Date _____

Acknowledgements

Living in Canberra, the quality of one's life depends on the quality of one's friends. I would like to thank in particular the following remarkable people for their friendship and support over the last few years: Wolfgang Schwarz, Leone Miller, Rhiannon Long-Rabern, Jens-Christian Bjerring, Weng Hong Tang, Clas Weber, and John Cusbert.

I have benefitted greatly from the intellectual environment in the RSSS Philosophy Department at the Australian National University, and would like to especially thank the following people: Geoff Brennan, Daniel Star, Jens-Christian Bjerring, Weng Hong Tang, John Cusbert, John Matthewson, Ole Koksvik, and in particular, Wolfgang Schwarz.

I am grateful to audiences at the University of Auckland, Waikato University, the Australian National University, Oxford University, Johann-Wolfgang-Goethe University Frankfurt, Humboldt University Berlin, University of Manchester, and the University of Concordia Montreal for helpful discussions and feedback.

I am indebted to Pablo Gilabert for many interesting exchanges about feasibility.

Finally, I am especially grateful to my supervisors Nic Southwood, Kim Sterelny, Lina Eriksson, and Brett Calcott, for reading various drafts of these chapters and for many interesting and challenging conversations, and above all to my primary supervisor Bob Goodin, for his invaluable insight, and prompt and always helpful feedback on the many stages these chapters have been through.

Abstract

This is a thesis about feasibility constraints for normative political theories. Political theorists talk about theories being 'utopian', 'ideal', 'abstracted', and more obviously pejoratively, 'impractical' and 'unrealistic'. They also talk about theories being 'non-ideal', 'practical' and 'realistic'. What exactly is at the heart of these charges? In the first chapter of this thesis, I discuss the emerging debate over ideal and non-ideal theory. I argue that the criticism of ideal theory is largely unjustified, but that it is important that theories intended to be action-guiding take feasibility constraints seriously. The main question I try to answer in the thesis is what exactly feasibility constraints for normative political theories consist in. I argue that there is a binary sense of feasibility (feasible, or not) that follows traditional discussions over the principle that 'ought implies can'. That is a useful sense, but it does not do all the work that feasibility constraints needs to do. Thus I defend a second, comparative sense of feasibility (more or less feasible). Most of the work of the thesis goes into elaborating these two senses of feasibility, establishing the kinds of facts that count against the feasibility of a proposal, and discussing the way feasibility assessments should be made in practice. In later chapters I consider whether those two senses of feasibility apply as well to collective agents as they do to individual agents. I also ask whether our cognitive and epistemic limitations affect our ability to make reliable feasibility assessments.

Contents

Chapter 1

Introduction: The Concept of Feasibility in Political Theory

> Even if a large dose of selfishness is part of human nature that does not refute theories of justice that require people to be less selfish than that (Estlund, 2010).

> How could this possibly *guide* anyone, other than to the madhouse? (Jackson & Pargetter, 1986, p. 242).

There were times, in the not-too-distant past, in which black people were kept as slaves, and in which a woman's worth was established only by her youth and appearance. At either of these junctures in history, we might have asked 'is emancipation feasible?' 'Is suffrage feasible?' Or, 'is it feasible to change the world so that all people are treated as equals, regardless of the colour of their skin, or their gender?' We know the answers to these questions because we have borne witness to the changes that establish their answers. If some state of affairs is actual, then *a fortiori* it is feasible. Slavery has been abolished, and women have the right to vote. There remain inequalities between persons on the basis of their colour and gender, but these inequalities are much less severe than they were at their historical worst.

Today we engage in practices of factory farming on a massive scale, which results in what is equivalent to the torture of a huge number of animals on a regular basis. And many countries have nuclear weapons, which means there is a constant possibility of nuclear warfare. We might ask about these states of affairs too, 'is abolishing factory farming feasible?' Or, 'is universal nuclear disarmament feasible?' We do not know the answers to these questions, but we can make more or less educated guesses.

This is a thesis about those more or less educated guesses. Political philosophers all too often criticize one another's theories on the grounds that they are 'infeasible', or synonymously, 'unrealistic', 'impractical', 'impossible', 'utopian', or 'idealistic'. As Juha Räikkä has pointed out, “the notion of political feasibility has always been politically significant. There has always been some tendency to reject new and unbiased suggestions on the grounds that they are “impossible” or “idealistic” or “utopian”. This is why it is practically important to be aware how the notion of feasibility is used and should be used in political arguments that represent political theory” (Räikkä, 1998, p. 39). I'm interested in what the criticism of 'infeasibility' actually amounts to, and the extent to which that criticism actually hurts a theory that is fairly accused of it.[1]

1Räikkä (1998) distinguishes 'political feasibility' from 'the feasibility condition in political theory', arguing that a theory can meet the feasibility condition without being *politically* feasible, by which he means that

A strong version of the criticism would be that there are some facts about how the world is that *every* political theory must take seriously. To borrow an example from Pablo Gilabert, a theory that requires all citizens to be provided with more than the average national income is infeasible because it violates a constraint that any theory must respect, namely the constraint of logical consistency (Gilabert, 2008). The average income is calculated by adding the incomes of each citizen within the nation, and dividing that sum by the number of citizens. No increase in income could make it so that *every* person earned *more* than the average; that is just not how averages work. If that strong version is the best version of the criticism, then the project is to figure out just what kinds of facts every political theory must take seriously, and what the effect of their not doing so would be. If a political theory violates the requirement of feasibility understood in this strong sense, is it *not a political theory at all*? Is it just a *bad* political theory? Or is it a totally acceptable political theory, just one that happens to be *infeasible* rather than *feasible*?

A weaker version of the criticism would be that there are some facts about the world that *some* political theories must take seriously. For example, theories that issue requirements for action make claims about what is obligatory, and the theorists who issue them are usually committed to the claim that people ought (in the practical reasoning sense, rather than the moral philosophy sense) to do what they are morally obliged to do. So if an animal liberation activist defends the requirement that people immediately stop eating factory-farmed meat, or a nuclear disarmament activist defends the requirement that countries immediately and multilaterally disarm, then it must be feasible that people stop eating factory farmed meat, and that countries disarm. If this weaker version is the best version of the criticism, then the project is to first figure out which kinds of theories are subject to feasibility requirements, and to then say what kinds of requirements these are.

At first glance, this project has a lot in common with the moral philosophers' discussions about the principle that 'ought implies can'. They argue, for example, that the proposition that *X* is obliged to φ is *false* if *X* cannot (on some suitable understanding, like 'lacks the ability to') φ. Their idea is that normative requirements are subject to an empirical constraint. By contraposition (which is permitted contingent our taking the

the political will to bring the theory's recommendations into practice might be lacking. I am concerned in this thesis with 'the feasibility condition', and try to talk in terms of 'feasibility constraints' or 'requirements of feasibility'. Any usage of 'political feasibility' should however be read as synonymous with those, rather than as indicating an interest in distinctively political constraints.

relevant implication to be entailment, but not otherwise), 'ought implies can' gives 'not-can implies not-ought', or, more colloquially, 'if it's not the case that a person can φ then it's not the case that she ought to φ'. The structure of the questions we might have about the principle that 'ought implies can' is the same as the structure of the questions we might have about the allegation that political theories ought to obey feasibility constraints. Namely, are *all* moral obligations subject to 'can' constraints? And if not all but only some, *which* kinds of obligations are subject to these constraints, and what are the constraints, exactly? The similarity of the discussions is pronounced, but it is important to note where they come apart. Moral philosophers are in general concerned with the *actions* of *individual* agents, almost always in the current temporal context. Political philosophers, in contrast, are in general concerned with *outcomes*, which can be brought about by either individuals or *collectives*, almost always over some extended temporal period. For example, moral philosophers might ask which action, available to a specific agent, in a specific context, is the action that ought to be chosen. Political philosophers might ask which outcome, not necessarily achievable by a set of actions available to any currently existing agent, would best realize the values she is committed to, and therefore which should be *chosen as a goal*, or *aimed at*, in the foreseeable future. We can draw from the moral philosophers' discussion, but we must be careful to notice these divergent concerns.

There is a commonsense, ordinary language notion of feasibility. It is something like 'can be realized in practice'. International nuclear disarmament is feasible if there's something we can do to make it happen. Animal liberation is feasible if there's something we can do to bring it about. I am going to defend something in the neighbourhood of that commonsense notion in this thesis. I will argue that much depends on how we specify the *subjects* for whom some action or outcome is feasible, and that different ways of specifying those subjects lead to different senses in which some outcome is feasible. Clarifying the ambiguity in the concept will permit less by way of theorists talking past each other. If one theorist criticizes another's theory on the grounds that it is infeasible, and a second theorist disagrees with that criticism, we can resolve the apparent disagreement by showing that the first theorist is using 'feasible' in one sense and the second theorist in another. What is important to establishing theoretical disagreement is whether the theorists disagree about whether a theory is feasible once they have settled upon a common understanding of *what it is* for a theory to be feasible.

Having the potential to resolve philosophical disagreement is one advantage of getting clear about the concept of feasibility in political theory. But there are other advantages. The concept of feasibility can play several important theoretical roles. It is a tool we can use to *rule out* theories (or a theory's recommendations) when they absolutely cannot be implemented. It is a tool we can use to *rank* alternative theories, to say that one is less practicable than another, to comment on the extent to which any given theory is practicable. It can be used to tell us about the *powers* that certain things have, to undertake particular actions, to bring about particular outcomes. It can also be used to *supplement practical reasoning* about action choice, and as a *heuristic* to decide which action to choose when the full decision-theoretic calculus is unavailable. Let me say a little more about each of these.

Feasibility can be used as a fairly blunt tool to *rule-out* political theories or recommendations which absolutely cannot be implemented. Imagine that, having split the atom, it is impossible to revert to a state of the world in which there are no nuclear weapons. Once invented, nuclear weapons and the possibility of nuclear warfare cannot be uninvented, in which case the nuclear-free future imagined by supporters of universal disarmament is not one that can be brought about. If a theory makes a recommendation that *cannot* be implemented in this fairly absolute sense, then it is *ruled out*. Ruled out of what? That is a question that remains to be answered, and will depend on whether we choose the strong or the weak version of feasibility. So far I have talked about both theories and theories' recommendations as being subject to feasibility constraints. Recommendations are direct, so they can more easily be ruled out by our showing that they cannot be realized in practice. But theories are just sets of principles that together (a) paint a picture of a just world and (b) make normative recommendations. If there is no way of making our world into the just world, then the ideal of the theory is infeasible. That is not in itself enough to warrant taking the theory out of consideration. But if in addition the normative recommendations of the theory, and all the manifold ways that these could be realized, are all unrealizable in practice, then it follows that the theory itself is infeasible. That is why both whole theories, and theories' recommendations, can be ruled out as infeasible. In this role, feasibility is binary. A theory is either feasible, or not.

In that first role, feasibility is a fairly blunt tool. It rules out theories or recommendations that absolutely cannot be implemented. In its second role, the tool can do more fine-grained work. It can allow us to say of theories or recommendations *how*

feasible they are. In this role feasibility is graded, rather than binary. We can make pairwise comparisons between any two theories, or recommendations, and we can rank a set of theories or recommendations in terms of how feasible they are, the extent to which they can each be realized in practice. For example, imagine that one animal liberation theorist proposes that we stop eating factory-farmed meat, and consume only meat and dairy products committed to the humane treatment of animals. And imagine that another demands that we become vegans, and liberate all animals currently kept in domesticity. A lot depends on the details of this case, but it is plausible at first glance at least to suppose that the recommendations of the former theorist are more feasible than the recommendations of the latter. This assessment rests on something like the commonsense assumption that the possible world where we start eating meat and dairy products committed to better practices is *closer* to the actual world than the world where we stop eating meat and dairy products all together (although I will prefer a different approach than 'closeness' for establishing feasibility in the thesis). If that assessment is true, we can say that universal veganism is *less feasible* than universal refusal to consume products from companies with inhumane practices. In this role, feasibility is graded. One theory is more or less feasible than another, or a theory is feasible to a greater or lesser degree.

Related to this comparative sense, the third role feasibility can play is in telling us about the *powers* of certain beings. Does North Korea have the power to send a man to the moon? Does Faber-Castell have the power to take over Crayola? Does Chiara have the power to quit smoking? We have intuitive ideas about the powers agents have, but sometimes we might want to say more than that they have them or they don't, but rather say something about the extent to which they have them. Thus there is a third role for feasibility, using the comparative role as a platform, in identifying the powers or abilities that particular agents have, be them animals, persons, companies, or nations.

Feasibility has its fourth and fifth roles to play in action choice, as formalized by both the causal and evidential versions of decision theory. Decision theory tells us to maximize expected subjective value. What does this entail? We calculate the value of an action by first thinking about all the possible outcomes of that action. We think about how desirable each such outcome is, and what the probability of the outcome coming about, given the action, would be. We multiply the desirability of the outcome by the probability of the outcome given the action, and we take that number and add it together with all other such numbers for all other possible outcomes of the action. That

sum is the expected subjective value of doing the action in question, and it can be compared with the subjective expected value of doing alternative actions. Where does feasibility come into this process?

It comes in first as part of the probability calculation. Saying how desirable an outcome is is a normative project, but assigning a probability to an outcome conditional upon an action is an empirical project. Feasibility in its ruling out role is a tool that lets us say when a recommendation cannot be realized in practice. And in its graded role it is a tool that allows us to say *how* feasible some recommendation is. Both of these roles are useful when it comes to thinking about how likely some outcome is, given a particular action. Decision theory requires that likelihood as an input, but it says nothing about how to obtain it. So getting clearer about feasibility, e.g. what it consists in, and what kinds of facts are relevant to determining it, will be helpful in making the probability assignments necessary to decision theory.

This is not to be confused with the idea that feasibility is a positive tool for action choice. It is a negative tool insofar as in its binary role it can rule actions out (although, if it rules out all actions but one, it obviously does make a positive recommendation). But in its comparative role, it cannot *alone* say which action in a given set of (ruled in) actions should be chosen. To make this positive recommendation, comparative feasibility assessments must combine with normative considerations. If the actions in a set are all feasible in the binary sense, and all *equally* desirable, then the most feasible should be chosen. But this will be unusual; and where it is not the case, there is no formal substitute for the use of judgement.

The second way in which feasibility comes into the decision theoretic process is as a decision-making heuristic. The decision theory formula is complicated. Take the global poverty problem, and suppose a person has exactly three actions available to her: (a) sponsor a child in Zimbabwe, (b) donate to Kiva (who facilitate micro-lending to small businesses in developing countries), and (c) join a local group of activists and spend time trying to convince others to support governmental policy reform with respect to international aid. Supposing she can only do one of these, the agent will be interested to work out which action has the highest value. But that entails thinking about all the possible outcomes of any one of the actions. It requires saying how desirable each of those possible outcomes is, somehow figuring out the probability of each of those outcomes given the action, then for each possible outcome multiplying the desirability by the probability, and finally adding all of those further sums together. She then has to do

all of that again for the other two actions, and finally can compare those value assignments to see which is greatest. Only at that final stage does she have an action that is all things considered choice-worthy. If the agent wants to maximize subjective expected value, she will choose to do the action with the highest value.

Now imagine that the action which turns out to maximize value is option (b), donating to Kiva. Suppose the reason this action comes out as best is that it minimizes risk. That is to say, for every donation the agent makes to Kiva, there is an extremely high probability that it will be paid back, and that it can then be used to make recurring loans to other businesses. A small amount of money goes a long way, and the risk of its being misused is negligible. Let's say the reason action (a) does not fare so well is that the risk of misappropriation is higher. When I sponsor a child, some percentage of my money goes to organizational costs and advertising, not directly to the child. And it has been often reported that these funds are misused. And (c) has a less high value because of the opportunity costs associated with the agent donating her time to the activists' campaign, in comparison with the other things she could be doing, including (b).

Decision theory takes account of risk in a way that feasibility alone does not. When we think how feasible an action is compared to other actions, we generally think about what chance it has of succeeding in practice. We focus on the *chance of success* rather than the *risks of failure*. So an action that has an 80% chance of succeeding will be seen as more feasible than an action that has a 70% chance of succeeding, even though the latter action might have a 30% chance of producing only a slightly less desirable state of affairs, while the former might have a 20% chance of producing a state of affairs that is catastrophic. That is not a problem with feasibility; that concept is not supposed to do the whole job of telling agents which actions to choose. Perhaps we cannot even say that an action is more choice-worthy just because it is more feasible. Feasibility counts for something, but so do risks, and so does desirability. We may well maximize expected subjective value by doing some action that is ranked lower in a graded feasibility set than several other actions.

But while feasibility cannot do the whole job of telling us what to choose, it may function as a heuristic. Most likely we will not have time to run the expected value calculus every time we have to make a choice over a set of actions. Then it is possible to do something rough and ready, like ask: which of these actions is the most feasible? And which is the most desirable? If we try to optimize with respect to both of these considerations, we will probably end up choosing options that are 'good enough', even if

not what the full decision theoretic calculus would have recommended. Certainly this heuristic is not perfect, and can get things wrong sometimes. For example, as mentioned already, we might choose an option that scores well with respect to feasibility and desirability, but which comes with some small chance of catastrophe. And risking that catastrophe might be worse than choosing a slightly less feasible and less desirable option which comes with no such risk.

The fact that feasibility plays a role in decision theory in these two crucial respects, first in contributing to assigning conditional probabilities to outcomes given actions, and second in acting as a heuristic when it wouldn't be prudent to run the full decision theoretic calculus, means that we have to be very careful in how we talk about feasibility and the work it does so as not to suggest that it's meant to *take the place of* decision theory. It is not meant to do that (and for the reasons discussed already, wouldn't do a very good job of it).

While I will sometimes talk about theories and their recommendations in what follows, I will also sometimes talk directly about obligations upon individuals to do certain actions. For example, I might say that consequentialism as a broad normative theory recommends that people minimize suffering (or maximize welfare, or preference-satisfaction, or happiness, depending on your preferred version of consequentialism), and then ask whether it is feasible for people to do that, and with reference to what kinds of facts we are to determine whether it is or isn't. But now take some recommendation that consequentialism makes in a given situation. Suppose an individual happens to be walking along when a car hits a cyclist and then speeds away without stopping. The moral imperative to minimize the amount of suffering in the world would dictate that the individual go over to help the cyclist; to check whether she is hurt, to help her and her bike off the road, to assist her in getting medical help if she needs it. (Most consequentialists would add a caveat like 'only if it does not seriously inconvenience her', which makes room for the possibility that she is a doctor who could minimize suffering better by going on to work). Feasibility considerations at no point tell the individual what she should choose, all things considered. It matters if she *can't* help the cyclist. It might matter to which outcomes are feasible for a third agent that she *won't* help the cyclist. But declaring that it is feasible that she help the cyclist, or that it is more feasible than some other options that are available to her, doesn't get us to a claim about what she ought, in the practical reasoning sense, to do (unless coupled with a true obligation statement).

So much for the five central roles that the concept of feasibility, in its different

forms, can play. Now it is time to start answering some of the more difficult questions posed early in this chapter, namely, are all theories subject to feasibility constraints? If not all, then which ones? And what kinds of constraints are they subject to? How do we establish that a theory is feasible, or more feasible than another, and are there any limits on our being able to establish those things? I shall proceed as follows. In Chapter 2 I introduce the debate between ideal and non-ideal theory. That is to address the question 'are all theories subject to feasibility constraints?' I will argue that most of the criticisms of ideal theory are unsuccessful, but that does not vindicate our exclusively *doing* ideal theory. I will argue that there is an important place for non-ideal theory, theory that aims to be action-guiding, or policy-shaping, and that feasibility constraints are most important to that kind of theory. Then for the remainder of the thesis I will concentrate on non-ideal theory.

In Chapter 3 I discuss some background to the concept of feasibility in political theory, in the first part of the chapter canvassing some of the ways that the concept is used, and in the second part of the chapter considering several aspects of the theory that have been defended in the literature. In Chapter 4 I introduce the discussion about 'ought implies can', which I borrow from as a way of developing and formulating a binary test of feasibility. In Chapter 5 I concentrate on developing the comparative test of feasibility. In Chapter 6 I talk about the role feasibility plays in practical judgement, expanding on the two roles suggested earlier that feasibility can play in supplementing decision theory. I ask what other considerations go into a judgement that some action is that which I ought to do, and I ask whether we are good enough at making probability assignments for any of this to be considered remotely practicable. In Chapter 7 I ask whether the comparative feasibility test is adequate to the task of describing the feasibility of collective action, arguing that it is, but that we must accept a certain account of how abilities distribute from individuals to groups for that story to work. In Chapter 8, I concentrate on the background conditions against which an agent's action takes place, asking, especially for the collective case, under which general conditions our actions will be more and less likely to succeed. Chapter 9 I spend concluding.

Chapter 2

Ideal and Non-Ideal Theory

> Nonideal theory must steer a course between a futile utopianism that is oblivious to the limitations of current international law and the formidable obstacles to moral progress erected by vested interests and naked power, on the one hand, and a craven capitulation to existing injustices that offers no direction for significant reform, on the other (Buchanan, 2004, p. 61).

2.1 Introduction

Within contemporary political philosophy there is considerable debate over the relation between ideal and non-ideal theory. The long-standard methodology has been to figure out principles of e.g. justice or equality by figuring out what these things would be like in an ideal world. Recent criticisms of that strategy focuses on the dissimilarities between an ideal world and our own, arguing that we can know everything about justice for an ideal world and still have no idea about justice for our own world. Roughly speaking, the counterproposal is that we should start with what we have and figure out how to make it better, not start with what is best and try to figure out how to apply it to what we have. The debate raises many issues. Is ideal theorizing as distinct from non-ideal theorizing as its opponents take it to be? If so, are they mutually exclusive, or only distinct as a matter of practice? Might they be complementary, different but necessary aspects of a larger enterprise? Might it be appropriate that some theories are ideal, and inappropriate that others are, depending on their subject matter, or their aims?[2]

I address those questions as a way of locating the feasibility project that is the concern of this thesis within the ideal/non-ideal spectrum. There is a prima facie tension between ideal theory and a concern with the politically feasible, given that in ordinary language the former suggests an extrapolation away from empirical facts, and the latter a serious concern with them. But this tension only arises if one thinks that *all* political theory should take feasibility constraints seriously. That would be to side with the non-ideal theorists against the ideal theorists. But I will argue that both ideal and non-ideal theory have a valuable role to play. I will show in the rest of this chapter that the debate over ideal and non-ideal theory constitutes the big picture inside which issues of political feasibility are located. Having sketched the logical space of that bigger

2 One important and by now often overlooked paper in this area is due to the economists Richard Lipsey and Kevin Lancaster (Lipsey & Lancaster, 1956-7). For a recent survey of the ideal theory literature see (Simmons, 2010).

picture, I shall then focus on the smaller topic of political feasibility in non-ideal theory.

In Section 2.2 I outline the challenges to ideal theory. In Section 2.3 I discuss some of the contemporary defenses of ideal theory against those challenges, and in Section 2.4 I conclude with a view on the place of political feasibility within the ideal and non-ideal theory discussion.

2.2 Ideal theory and its enemies

In June 2010 an online protest began, under the heading 'Stop Female Genital Mutilation at Cornell'. A group of individuals were angry at what they perceived as the mutilation of the genitals of children by a surgeon at Cornell University in Ithaca, New York. The surgeries were not, as one might have expected, instances of clitoridectomy, but rather gender reassignments in intersex children. This is a good case for introducing the distinction between ideal and non-ideal theory. The Intersex Society of North America (ISNA) estimate that around 1% of live births exhibit some degree of sexual ambiguity, between 0.1% and 0.2% enough so that medical attention is justified (ISNA, 2010). But in most societies, there is little to no awareness of this fact. Gender assignments are clear cut: there are males, and there are females. Furthermore, given the strong gender roles that characterize most modern societies, it is likely that an intersex child entering adolescence might end up suffering a lot; with doubts and insecurities about their own sexuality, and at the mercy of rejection and alienation from peers and prospective sexual partners. That is not to say that such suffering is *inevitable* for an intersex child, but that ambiguous genders are not commonplace in the mainstream, and so can be expected to meet a certain amount of resistance. Those are the characteristics of the non-ideal world we happen to be in. Now *given* those characteristics of the non-ideal world, there is a genuine question about what a parent should do when faced with the choice of requesting surgery for an intersex child, or not.

One of the costs is assigning the child a gender he or she may later fail to identify with (it is easier to reassign intersex children to female genitalia than male, so surgeries are more commonly in that direction). The advantages are potentially avoiding a lot of the suffering and confusion that the child would otherwise experience growing up. It is at least understandable that a parent might opt for surgery rather than allow the child to decide for his or herself later in life, especially seeing as a social gender will have to be selected early on even if the physical gender doesn't completely coincide with it. One

way of approaching the issue of how a parent should decide is to take into account all the facts about our non-ideal world, and then think about what would be best for the child, all things considered. We can still admit that it would be better if the surgery were not necessary, if people were more accepting of ambiguities in gender. But given that they are not, certain choices will be justifiable.

The complaint that non-ideal theorists make against ideal theory, as we shall soon see, is that it has nothing to say about this kind of case. Maybe in an ideal world, children are not born with any gender ambiguity. Or maybe they are, but people are much better educated about the possible range of genitalia, and much more accepting of individuals who fall between the straightforwardly 'male' and 'female' categories. In that kind of world, surgery wouldn't be necessary. Does that mean that surgery is not necessary? Obviously not, because we do not find ourselves in an ideal world. This is what grounds the complaint that ideal theory gives no guidance as to what should be done in the real world.

We can see from this case that there is a sense in which that complaint is true, and a sense in which it's false. It's true that knowing what the ideal world is like doesn't give the parent any guidance on whether to choose surgery for her child in the non-ideal conditions she finds herself in. But it's not true that knowing what the ideal world is like gives no guidance whatsoever. If we agree that better education would render such surgeries unnecessary, we can try to educate people more in the non-ideal world we are in, to try to make the non-ideal world more like the ideal world in the relevant respects.

In the rest of this section, I shall present the most notable attacks on ideal theory. I begin with the most common attack, which along the lines just described is that ideal theory fails to be action-guiding. This complaint takes various forms. For example, Onora O'Neill claims that theory is pernicious when it reasons from false premises; Laura Valentini claims that it is pernicious when it builds in unmodifiable assumptions; Colin Farrelly claims that no conclusions from ideal theoretic premises follow for the non-ideal world; and Geoffrey Brennan and Philip Pettit protest against the hopelessness of assumptions about strict compliance. I also discuss an attack against ideal theory along methodological lines, due to Amartya Sen, and the complaint that ideal theorizing smuggles ideology into political theory, due to Charles Mills.

2.2.1 The standard line: ideal theory fails to be action-guiding[3]

Onora O'Neill (1987) notes that ethics has often been attacked for being too abstract. She thinks the attack is surprising, given that abstractions are unavoidable (no use of language is fully determinate), often admired (e.g. in law, and accounting), and allow the wide scope necessary for application in a range of circumstances (O'Neill, 1987, p. 55). She claims that one of the attacks on abstraction is motivated by conflating abstraction with idealization. Idealization involves omitting too much that we know to be true, and adding too much that we know to be false, to our assumptions about e.g. human agency. Doing so means we end up talking about hypothetical agents rather than real people. And worse still, we sometimes treat these hypothetical agents as *normative ideals*, standards which we should aspire to (O'Neill, 1987, p. 56). O'Neill agrees that theory based on idealization is objectionable (she basically takes this for granted), and gives examples of several theoretical attempts to modify prominent idealizations so that they are more reflective of real world agents and circumstances. But, she claims, abstraction is not idealization, and there's no reason that this methodology should be assumed objectionable merely by association.

O'Neill discusses several possible objections to abstraction, concluding that even though none of them seem to work, it is unlikely that the persistent intellectual current against abstraction in ethics will turn out to be groundless. She suggests that the main dissatisfaction with abstract theory might be that it says too little in general about deliberation, about how mutual comprehension, understanding and agreement with respect to the general premises of a theory, and application of its conclusion, can be achieved. If a theory says more on that point, or if it is possible for later theorists to say more on that point, the general worry about the theory being too abstract should be dissolved.

O'Neill (2000) returns to the abstraction / idealization distinction, and casts it in slightly different terms. Here she describes abstraction as the bracketing of certain truths, and idealization as reasoning from false premises. Abstraction is unobjectionable, because we're only bracketing; idealizing is objectionable, because we're literally asserting the absence of predicates that actually obtain, or asserting the presence of predicates that do not obtain. We should hardly expect the conclusions that follow from false premises to themselves be true (O'Neill, 2000).

3 Parts of this section draw on my (Lawford-Smith, 2010).

With respect to her (2000) way of casting the distinction, the categories don't seem particularly obvious. What is 'bracketing' if not asserting, for the purposes of the argument, that certain predicates do, or do not, obtain? Should we say that Newton's famous 'frictionless plane' thought-experiment in physics 'brackets' the existence of friction, and so is an abstraction, or 'asserts the absence of friction', and so is an 'idealization'? If a theory can be modified so that it reintroduces some of the material that was asserted absent (or bracketed) does that entail that the theory was only an abstraction? Or can theories be complicated and modified in either case? With respect to the (1987) distinction, it is clear that both abstraction and idealization correlate with ideal theory, but the distinction suggests a way to argue that only certain kinds of ideal theory are objectionable, namely those that idealize (reason from false premises, omit too much that is true, or add too much that is not). If the objection goes through, the attack on ideal theory is partially vindicated, because *some* kinds of ideal theory are objectionable. But it isn't fully vindicated, because there are still perfectly acceptable kinds of ideal theorizing.

Laura Valentini (2009) makes a similar distinction to O'Neill's, between 'good kinds' and 'bad kinds' of ideal theory. She concentrates on undermining the premise 'any ideal theory fails to be action-guiding', which is the premise in an argument for ideal theory being paradoxical. In general she argues in support of ideal theory, by saying roughly that not *all* ideal theory fails to be action-guiding. But her claim that some ideal theory fails to be action-guiding is interesting for the purposes of this section, for the reason that it seems to extend O'Neill's criticism (although it's not exactly clear whether Valentini's 'good' and 'bad' kinds of ideal theory map onto O'Neill's 'abstractions' and 'idealizations' respectively, or whether she is making a distinction within the category of 'idealizations').

Her argument is that bad kinds of idealizations idealize their *subjects* rather than mere background conditions, a kind of bracketing or assertion of the absence of some property that *can't be reintroduced later*. To apply this to the example just given, Newton's thought-experiment is acceptable, because although it assumes away friction for the purpose of finding out how bodies would otherwise behave, friction can be introduced later on. Imagine that the thought-experiment justified the prediction 'a moving body without external obstacle and in the absence of friction would stay moving', and that we could from that prediction, knowing the general effects of friction, make the further prediction that 'in these conditions there is a lot of friction, so the moving body should

slow down fairly rapidly and then come to a complete stop'. Valentini's idea is that bad idealizations would not allow such a reintroduction. She thinks that economists' assumptions that people are perfectly rational are an example of this. Economists have assumed that people are perfectly rational value maximizers, and made predictions as to how they would (or should) behave. But Valentini thinks there is no point at which economic theory can 'build back in' the fact that people are not perfectly rational. The idealization was about the subject of the theory, not just the background conditions. And thus the idealization tells us only about what would happen *if* people were as ideally hypothesized, not that there is some way people are absent other conditions, in a way that we can slowly reintroduce the bracketed conditions to get a picture of how imperfect rationality will be once those conditions are accounted for.

One final example of criticism of ideal theory along this standard line comes from Geoffrey Brennan and Philip Pettit's understanding of ideal theory as theory that assumes strict compliance (which is a narrower understanding than we have been considering so far, but a common one thanks to Rawls (1971)). They object that this assumption does not reflect the conditions of the real world: 'almost any set of principles for the organization of society, and certainly any principles of justice, are going to be burdensome for its members and so are not going to attract universal compliance' (Brennan & Pettit, 2005, p. 260). They argue that normative theory's failure to take compliance constraints seriously threatens to undermine the whole normative enterprise. One reason they give against ideal theory is the problem of the second best, (originally due to (Lipsey & Lancaster, 1956-7), but see also (Brennan, 1993) and (Goodin, 1995; forthcoming)). This is the problem that once any one of the conditions for Pareto-optimality cannot be fulfilled, the others become suboptimal, so that the only way to achieve optimality is to depart from them all (Lipsey & Lancaster, 1956-7, p. 11). Or more informally, approximations to the best state of affairs might not themselves be desirable, often because of necessary tradeoffs.

Brennan and Pettit argue that normative theory should buy into 'incentive-compatibility', familiar in Economics, where normative arrangements are 'compatible with incentives [...] that routinely and reliably affect what people do'. They argue that arrangements must be of a kind that 'ordinary human beings are able in general to sustain', and must not be 'motivationally too demanding' (Brennan & Pettit, 2005, p. 264). In fact, they point out, this kind of non-ideal theory is actually historically familiar, in the neo-Roman republicanism that was to influence Machiavelli, Harrington,

Montesquieu, and the writers of the Federalist Papers (Brennan & Pettit, 2005, p. 264). But it has been largely lost since Rawls (1971) (which is also a reason to resist the presentation of the ideal theory debate in (Simmons, 2010)).

They suggest that one kind of incentive we might try to ensure that political reform is compatible with is esteem-based. That avoids the problem that material incentives can actually be counterproductive, working against rather than alongside people's naturally virtuous motivations. The problem is to find the right balance between taking people to be knaves, and taking them to be perfectly virtuous. Too much of either yields a theory that is likely to make counterproductive recommendations. If we try to make political reform compatible with esteem, then we can channel people's ordinary virtuous motivations, while also giving them further incentives to comply (e.g. greater esteem and the positive effects that come along with it) without those incentives (e.g. the material incentives just mentioned) working against ordinary motivations (Brennan & Pettit, 2005).

To recount, O'Neill is concerned about reasoning from false premises, omitting too much that is true, or adding too much that is false, to the premises of arguments. Valentini is concerned about arguments that do not allow the reintroduction, at a later stage, of what has been assumed away, and Brennan and Pettit are concerned about the assumption of strict compliance generating counterproductive normative recommendations. While we can agree with Brennan and Pettit that it is important to find the right balance between treating people as angels and treating them as knaves, I'm not convinced that any of these alleged failings of ideal theory carry as much weight as their authors believe. What does it matter if we model human agents as perfect reasoners whose abilities are impeded by other conditions *making* them imperfect reasoners, instead of imperfect reasoners from the start? And what does it matter if we model moving bodies as in perpetual motion but impeded by external obstacles including friction, instead of as bodies impeded in various ways from the start? So long as we take account of the 'imperfect' conditions at some point between theory-design and implementation, the idealization/abstraction, good/bad ideal theory distinctions don't seem particularly useful. Maybe their claim is supposed to be that the bad kinds (the idealizations) *can't* handle taking account of imperfect conditions. But what reason do we have to accept that claim? One reason to reject it is that in many cases, even the bad kinds of ideal theory (idealizations) can *relax* their particular idealizing assumptions in a way that makes them applicable to the real world. We can estimate the distance

between a given ideal and the actual, and relax the stringency of ideal requirements in a way appropriate to actual circumstances.

We won't be able to relax ideal assumptions in all cases. For example, returning to the economists' assumption of perfect rationality, imagine that an ideally rational agent can perform infinitely many steps of reasoning in a given system of logic. And imagine that an actual agent can perform arbitrarily many finite steps of reasoning in the same system. In terms of 'approximating' ideals, it might be the case that one actual agent can perform more steps than another actual agent, and so in that sense is closer to the ideal. But in another sense, both agents' capacities are finite, so they are both infinitely far away from the ideal,[4] and the ideal is impossible for either of them to attain.[5] But why not think that the ideal can be relaxed to take account of the capacities of actual agents? Why can't we say something like 'well, ideally you would perform infinitely many steps of reasoning, but I suppose you have other things to do, so I will expect you to perform roughly 12, which you can do if you try'? Or we might build expectations out of statistical averages. Imagine that we expect rational agents to act in accordance with their preferences. And imagine that we happen to know roughly what their preferences are. Then we can make predictions about how they should behave. But knowing that real humans are only imperfectly rational, we can modify our predictions to allow for normal human error – changes of mind, irrationality, weakness of will, impulsiveness, and so on. If we know a bit about what people are like *and* we have an idea of the ideal, it is not impossible to 'expect less', informed by but not constrained by both the ideal and the real world.

If we're concerned with the gap between the Rawlsian ideal (strict compliance) and the non-ideal (imperfect compliance), then this idea of 'expecting less' is vulnerable to Liam Murphy's (2000) argument that we need a bridge from ideal theory to non-ideal theory, a way to translate ideal obligations into non-ideal obligations. If we don't have a *way* of translating e.g. the requirements of ideal rationality into requirements of non-ideal rationality then any claims about deriving the latter from the former will be empty. Murphy (2000) provides such a bridge, which shows at least that the translation is conceptually possible. His claim, with respect to non-ideal duties, is that a non-ideal duty is what would have to be done by the individual if she and everyone else were to do as

4 In other words, assuming that idealized agents in this example have infinite capacities, anything less than infinite falls infinitely short of the ideal.

5 I owe this example to Jens-Christian Bjerring, and am grateful to him for helpful comments on this section of the chapter.

they should. To try to flesh this out with an example, consider some ideal state of affairs, e.g. there being no poverty. If everyone were to do as they should, then (let's assume) everyone would contribute some fraction of their income on a yearly basis, which could be used in aid and in building infrastructure and institutions in developing countries. This ideal duty is divisible between all those who could, assuming strict compliance, fulfill it. So imagine it would take 2% of the gross domestic product of developed countries to realize the ideal, and that amount was divisible into *X* dollars a year for every normally-functioning adult in a developed country. Each person's non-ideal duty is to contribute that amount toward ending global poverty. So non-ideal duties fall out rather straightforwardly from ideal duties.

But Murphy's bridge is not without its problems. Sometimes it will be too costly to do one's share when no one else is doing theirs (it puts those who comply at a comparative disadvantage), and sometimes there will be thresholds where there is only value in a certain proportion of people doing their share, and no value in people doing their share below that threshold. However, the bridge is only meant to translate an ideal duty into a non-ideal one. It's not necessarily the case that the non-ideal duty it produces has to be what a given agent has, all things considered, most reason to do. I mention Murphy's attempt to bridge the gap between ideal and non-ideal theory only by way of noting that if it turns out to be too difficult to tell a convincing story about *how* ideals can be relaxed, then my claim that even the supposedly 'bad' kinds of ideal theory can end up being action-guiding won't go through. But that is not a big problem. So long as ideal theory is not ruled out in its entirety, the basic big picture view I want to defend later in the chapter remains intact. And nothing we have seen so far suggests that *no* kind of ideal theory is theoretically permissible.

Colin Farrelly is even more outspoken in his criticism of ideal theory, attacking it as a whole as being 'ineffective', 'failed normative theory', that yields 'impotent prescriptions', and collapses into idealization in O'Neill's sense (Farrelly, 2007, pp. 845-55). He attacks Rawlsian and Dworkinian ideal theorizing, both of which he classes as only moderately ideal because the theories do take some facts, like pluralism, seriously. The idea seems to be that if he can show moderate ideal theory to be objectionable, stronger ideal theory will be objectionable *a fortiori*. He argues against Rawls, for example, that his theory of justice fails to take scarcity seriously. That means the theory 'yields impotent prescriptions for real societies that face conditions of scarcity' (p. 848). The same accusation is repeated later, when he says that citizens in Rawls's Original

Position are supposed to think about their likely place in a society where the circumstances of justice are present, and under reasonably favourable conditions (p. 849). But, Farrelly complains, 'we do not know how rich or poor our society will [actually] be. So Rawls is not justified in claiming that the parties can assume that whatever society they end up in ... will be one in which the reasonably favourable conditions that make a constitutional democracy possible exist' (p. 850). To put this in the language of O'Neill's distinction, Farrelly's claim is that Rawls reasons from the false premises that (a) there is moderate or no scarcity, and (b) people will end up in a society where conditions are reasonably favourable, and where the circumstances of justice hold, to various conclusions about what justice is, or requires. But the premises are false, so *nothing follows* for the real world.

Farrelly's accusation involves a clear mistake. We can see it if we direct our attention to the obvious distinction between reasoning from a counterfactual conditional, and assuming the truth of its antecedent. Both Rawls (1971) and Dworkin (2000) can be understood as using a particular mechanism to obtain an idea of a just world. Both engage a hypothetical scenario: for Rawls, it is the original position with its veil of ignorance; for Dworkin, it is the desert island with its lack of prejudice and equality of bargaining power. These are thought-experiments, and they are common methodology in philosophy. Thought experiments are in most cases equivalent to counterfactuals (although there are some exceptions for e.g. rigid designators, see Jackson, 2003). For instance, while Rawls formulates the original position as a thought-experiment, we can easily reformulate it as a counterfactual: if citizens *had been* in a situation such that they lacked certain character traits, e.g. were not risk-loving, irrational, motivated by envy, and so on (see Rawls, 1971, p. 143; and discussion in O'Neill, 2000, p. 72), they *would have* chosen certain principles of justice to govern their political institutions, namely the difference principle and the equality principle. We can do the same for Dworkin's desert island scenario: if individuals *had* found themselves on a desert island, absent of prejudice, and so on, then they *would have* chosen a certain distribution of resources, and agreed to a certain scheme of insuring against bad luck.

Counterfactuals such as 'Rhiannon would have told Brian the truth, had he asked her how much she had spent shopping that day' are standardly understood to be properties of actual individuals. That means they are true or false. The counterfactual just mentioned is true if Rhiannon has the property of being such that she would have lied to Brian, had he asked her about her shopping expenses. Farrelly's mistake is that he

does not bother to separate the truth of counterfactual conditionals from the truth of their antecedents. 'If *p* □–> *q*' (following the notation in (Lewis, 1973) for 'if it had been the case that *p*, then it would be the case that *q*') might be true; to use Rawls's example again it might well be true that *if* we were not blinded by envy, risk-taking, and imperfect rationality, we'd choose certain principles of justice. But asserting the truth of the counterfactual conditional is not the same as asserting the truth of its antecedent, e.g. saying that it *is* that case that we're not blinded by envy, not risk-takers, not imperfectly rational. Certainly Rawls did not assert something so plainly false.

Consider an analogue. If orchestras had conducted their auditions blind (e.g. with those auditioning obscured from the view of those selecting) for the last century, there would have been a greater proportion of women in the world's best orchestras over that time (see discussion in Gladwell, 2005). The antecedent of the counterfactual is false, because orchestras did not begin the practice of blind auditioning until fairly recently. But it seems remarkably likely that the conditional itself is true, because the proportion of women in orchestras has increased dramatically since the introduction of blind auditions. And something normative is suggested by the truth of that conditional, namely that blind auditioning is something we should be committed to (given certain background assumptions about the value of gender equality). It's just the same with the counterfactuals in Rawls's and Dworkin's works. If their counterfactual conditionals are true – *if* we had been in a position to choose principles of justice, then...; *if* we had been on the desert island, then... – they have some normative force. If people would choose certain principles of justice were they not made blind and selfish by their psychological limitations, then perhaps those principles of justice are exactly the ones we should have governing our institutions.

What I want to suggest is that the attack on ideal theory put forward by e.g. Farrelly and O'Neill is an attack on the antecedents of ideal theoretic counterfactuals. With respect to the orchestra example above, it is like them yelling 'yes, but we don't have blind auditions, so nothing follows for real people!' That criticism is misguided because it focuses on the antecedents of conditional claims instead of focusing on the whole conditional. What's actually important is (a) whether the counterfactual is true, (b) what it recommends, normatively speaking, and (c) whether what it recommends is accessible. The work done by the counterfactual is to *justify* a good world. In the counterfactual world, people's reasoning about desirable principles of justice isn't marred by their psychological limitations, and they choose certain principles of justice.

That is how we get a picture of the good world, i.e. the world in which those principles of justice actually govern institutions.

There are three worlds, and two relations, in this picture: the actual world, the world where the antecedent of the counterfactual is true (the counterfactual world), and the good world justified by the world where the antecedent of the counterfactual is true. The important relation is not that between the actual and the counterfactual world, but rather between the actual world and the good world. Whether our actual world is anything like the world in which the antecedent of the counterfactual is true is beside the point. A much more interesting and potentially damaging criticism of an ideal theory would be that the ideals painted by those theories (in the non-technical sense of 'world we should aim for', see e.g. Valentini, 2009) are impossible to access from the real world. In the terms just discussed, that would be to say that the good world is not one we can get to from the actual world. *That* kind of objection is relevant, but for all Farrelly has said, the good world suggested by Rawls (i.e. one where the principles of justice govern institutions) might be accessible from the real world.

It is illuminating to consider Farrelly's idea of what justice is (he conducts his attack on ideal theory in terms of theories of justice): 'I believe there is some conceptual incoherence in saying "this is what justice involves, but there is no way it could be implemented"'. He thinks that both theories, and the principles they endorse, 'must function as an adequate *guide* for our collective action' (Farrelly, 2007, p. 845, his emphasis). It is mainly an artifact of this understanding of justice that his argument against ideal theory looks even superficially plausible. Anything that fails as an immediate guide for collective action will for that reason not even be a theory of justice. What Farrelly's view seems to come down to, as Zofia Stemplowska notices, is a mapping from 'ideal theory' and 'non-ideal theory' to 'useful' and 'useless' respectively (Stemplowska, 2008). But as we will see in Section 2.3 when we discuss ideal theory and its defenders, there is plenty to be said about the value of theory which is not directly action-guiding.

2.2.2 A methodological critique of ideal theorizing

Amartya Sen (2006) distinguishes the 'transcendental' from the 'comparative' approach to justice. Transcendental approaches specify the ideal of a perfectly just society, while comparative approaches focus on ranking alternative societies in terms of

their being more or less just. Transcendental approaches are ideal, comparative approaches are non-ideal. Sen argues that the transcendental approach to justice is neither necessary nor sufficient to the comparative. This would not be particularly objectionable in itself; what is problematic is his assumption that if ideal theory is neither necessary nor sufficient to non-ideal theory, then it is redundant. Sen thinks we should get on with things using the comparative approach.

He makes two claims which I think are mistaken. The first is simply that if ideal theory is neither necessary nor sufficient to non-ideal theory, then it is redundant. That is to assume that there's nothing else for ideal theory to do, than to work in the service of non-ideal theory. But in fact, one very important job of ideal theory is to do *conceptual analysis*.[6] It is common methodology in philosophy in general to appeal to far-fetched thought-experiments to find out about the nature of things. For example, most people were satisfied with the analysis of knowledge as justified true belief, until Gettier came along with cases where justification, truth and belief were present only accidentally, so that most people had the intuition that there was not knowledge (Gettier, 1963). Likewise many were satisfied with a materialist analysis of the mind until Frank Jackson presented the Mary case (Jackson, 1982) and Chalmers presented the zombie case (Chalmers, 1996), eliciting dualist intuitions. I give these examples by way of illustrating the point that one of the major jobs of philosophy is conceptual analysis, and there is little reason to think that normative theory should be any exception. So part of the point of thinking about perfectly just societies is to get a clear understanding of what justice (or fairness, or any other of the important normative political concepts) actually *is*. That is to defend the idea that there is something interesting and important about what Sen calls 'transcendental' and we call 'ideal' theory, *whether or not* the ideal world is accessible from our own. (See also Section 2.3.1 for Swift's distinction between different functions of political philosophy, which works equally well against Sen and accommodates the point I make here).

6 This assumes that normative concepts describe normative reality. But there are alternative views, e.g. in the constructivist family, that think normative concepts are solutions to practical problems (e.g. Korsgaard, 2003). On the former view, conceptual analysis is an epistemological exercise. On the alternative view, conceptual analysis (conceptual construction) is practical- different conceptions of social justice will consist in different solutions to the practical problem. Thus there is a prima facie case that the conceptual analysis defence of ideal theory, as I have posed it, begs the question against alternative understandings of normative concepts. (I am grateful to Laura Valentini for discussion on this point). To this objection I would simply say that there is a fact of the matter about what the best solution to a practical problem is- there is something true that can be said, describing normative reality, about the best solution to an unjust world, e.g. the perfectly just world. So the constructivist line collapses into the epistemological line, for that purpose.

The second claim I think is mistaken is that ideal theory is not necessary, because we can make pairwise comparisons between societies, saying which is more just, without necessarily having to know what a perfectly just society would look like. Sen uses two examples to make this point. The first is to aesthetics. He says that in arguing over which of a Picasso and a Dali is the better painting we don't have to have an idea of the perfect painting in mind. But this is to confuse a (wildly) subjective ranking with an objective ranking. For this to be convincing, Sen would need to argue that justice is like aesthetics. He acknowledges the possibility of this response (Sen, 2006, p. 222) and gives another example: we can compare the heights of the mountains Kanchenjunga and Mont Blanc without making any reference to the highest mountain, Everest (Sen, 2006, p. 222). Again, this example is unconvincing. There is no such thing as an ideally tall mountain; Everest is not an ideal, it just happens to be the tallest mountain that actually exists. Furthermore, it is not the concept 'mountain' that is doing the work in allowing the comparison between the Kanchenjunga and Mont Blanc, it is the concept 'height'. And height is a linear value, which uses a standard measurement (centimetres and metres). Certainly we do not need the concept of an ideally high thing in order to make pairwise comparisons between things in terms of their height. But in the other direction, it is not the case that knowing the ideal (when there is one) is *useless* in informing pairwise comparisons, as Sen tries to argue. Imagine that a person wants to sell her home at auction, and she tells the real estate agent that her ideal selling price would be $500,000. Monetary value, like height, is linear, so we know that in a comparison between any two offers, the higher offer is better, because closer to the ideal, than the lower offer.

Consider a concept which is purely relational, like 'being to the West of', and consider a concept which is monadic, like 'symmetrical'. For any two objects, we can say which is to the West of the other. Perhaps we come to learn relational concepts by being given many examples, from which we eventually extrapolate some kind of non-analytic definition. But there is no such thing as 'perfect Westness' (at least, not on a spherical surface like the globe). On the other hand, from any set of objects, we can say which is the most symmetrical. We have an idea of perfect symmetry (which can be realized in many and varied ways) and we can say based on this ideal which objects are closer to it and which are further away. That is just to illustrate that some concepts, like symmetry, require a fully-specified ideal, and some concepts, like Westness, do not. The question is, is justice more like symmetry, or more like Westness?

Sen obviously believes that it is more like Westness, but unfortunately he says

nothing – at least not in his (2006) – in support of that claim. Without saying something about it, he is vulnerable to the question of how a judgement that one society is better than another in terms of justice is made. And perhaps this is just the influence of the dominant Rawlsian paradigm that Sen is so valiantly trying to free his readers from, but it seems to me that what *justifies* judging one society more just than another society is precisely *having an idea of what justice is* (where knowing the concept comes from knowing the ideal, i.e. from doing ideal theory). For Sen's argument that we can make comparisons of 'more and less just' without having an idea of perfect justice to succeed, he must explain what makes those comparisons possible. Is it just like with Westness, where we learn from a handful of comparative examples? Of course we can make comparisons in terms of subjective preferences, but justice is supposed to be something more than the subjective preferences of one individual. At most it is something objective, at least it is something that reasonable people can agree about (even if they might disagree about some of the details). Until Sen gives a story about how comparative judgements are justified, I am inclined to think that we do need ideal theory in order to do non-ideal theory, because it is a good understanding of the normative political concepts that allows us to make justifiable choices between alternative political scenarios.

In summary, Sen argues that ideal theory is neither necessary nor sufficient to non-ideal theory, and is for that reason redundant and should be given up in favour of non-ideal theory by way of pairwise comparisons. Against him, I would argue that ideal theory is not redundant because it has a different role to play than merely to inform and assist non-ideal theory, and that it is in fact necessary to non-ideal theory, because without it nothing justifies choosing one alternative over another in a pairwise comparison. Sen's distinction between transcendental and comparative approaches may be slightly more illuminating than the labels 'ideal theory' and 'non-ideal theory' (at least they avoid the 'different kinds of idealizing' arguments of e.g. O'Neill 1987; 2000), but his arguments against the former do not warrant his conclusion.

2.2.3 Against ideal theory on ideological grounds

Charles Mills (2005) argues that ideal theory is ideological in the sense that it protects the interests of white, middle-class, men. He thinks that a focus on the ideal entails the exclusion or marginalization of the non-ideal. By ignoring the fact that in the actual world some are oppressed or subordinated by others, ideal theory ends up

reflecting and perpetuating illicit group privilege (Mills, 2005, p. 166). Mills makes it clear that by 'ideal' he's not concerned with theories that feature ideals (which all normative theory does), or theories that model some real system by selecting only essential properties of it. By 'ideal' he means to target *exemplars*, i.e. best versions of some thing, theories that model some real system by saying how it should function when working properly. He acknowledges that in addition to moral exemplars there will be functionalist exemplars (a properly working vacuum cleaner, an ideally functioning concentration camp), and exemplars utilizing limiting assumptions (frictionless planes, ideal gases) (Mills, 2005, p. 167).

He poses the question 'in trying to understand the workings of an actual P, how useful will it be to start from an ideal-as-idealized-model of P?' (Mills, 2005, p. 167). He thinks the answer will depend on how closely an actual token approximates the ideal model. If the two are very different, then we will need to theorize not only from the ideal model but also from the non-ideal token, to understand what prevents the token from attaining ideality (Mills, 2005, p. 167). He is right about this. If we decide it is desirable to move from the actual state of affairs to the ideal state of affairs, we will have to theorize about both in order to make the transition. What are the features of the ideal that we want to instantiate? What are the features of the non-ideal that will need to be transformed or eliminated? But it is not clear why this is a criticism of ideal theory. I cannot see that anyone doing ideal theory would deny that in moving from the actual world to the ideal world we will need to theorize about the actual world (although I would question the extent to which any ideal theorist actually thinks a direct transition from the actual world to the ideal world is possible).[7]

In a long list of complaints against ideal theory, Mills says that it has an idealized social ontology, it idealizes people's capacities, it is silent about oppression, it idealizes social institutions and the cognitive sphere (ignoring the way that experience shapes perception), and assumes strict compliance (Mills, 2005, pp. 168-169). Mills' challenge is 'how in God's name could anybody think that this is the appropriate way to do ethics?' (Mills, 2005, p. 169). He contends that middle class white males, who are hugely over-represented in the philosophical population, have an experience of reality which actually closely approximates their idealizations, which is why they do not experience the

7 Maybe a better criticism is that some ideals *underdetermine* the non-ideal stages of the world necessary to bring the ideal about. If there are many and varied ways to realize a given ideal (or if it is really difficult to judge when a stage is 'closer' to the ideal) then the ideal is somehow less useful than it would be were it to simply follow from the ideal what its realization would entail.

cognitive dissonance between ideal and non-ideal theory that marginalized people do.

The criticism that ideal theory 'ignores' injustice is baffling. If certain aspects of society are marginalized or excluded simply because they are not present in a theory, then ideal theory does 'marginalize' and 'exclude' all the non-ideal aspects of society. But *that's what makes it ideal*. By imagining a society governed by just institutions, where everyone has an equal opportunity to live a good life (and so on), we imagine away present injustices in favour of a world that is morally and politically ideal. But it's not clear why we should think that this counts as 'ignoring' non-ideal aspects of society. In fact, it obviously requires paying close attention to non-ideal aspects, recognizing that they are unjust or undesirable, so as to deliberately remove them from the ideal picture. Ideal theory recognizes all that is unjust or undesirable in the actual world and imagines it away. If a theory builds in oppression and subordination it is not the best of all possible worlds; if it is not the best of all possible worlds then it may be that we're taking some kinds of injustice as permanent features of the political landscape rather than targeting them for change. In the ideal world there are no people in terrible poverty, or without medical care, and women are not oppressed or subordinated, and neither are minority groups. That's not because all those groups are just swept under the carpet, it's because the systemic causes of all that injustice and unhappiness have been addressed. Poor people are not left out of a theory of ideal justice; rather they have become one of the standard subjects of it. The only justification I can see for Mills' claim is that he thinks certain relations of domination or oppression are unbreakable, in which case an ideal theory should make explicit reference to those who are 'essentially' worse off. But I think no one group of persons is *essentially* worse off.

A partially ideal world might contain special measures to remedy entrenched injustice (e.g. Susan Moller Okin's example that women are more constrained with respect to exit options from a marriage, which affects their chances in the job market, and which to be remedied would require special compensation (Mills, 2005, p. 178; Okin, 1989)), but a fully ideal world would be one where the job market no longer disadvantages women in the way it does in the actual world (it is a world where the sources of injustice have been remedied, rather than where special compensation is given after the sources of injustice have had an effect).

It seems to me that all this comes down to a choice between two ways of theorizing about justice. The first is to try to conceptualize *all* the worlds more just than our own, including the ideally just world(s – perhaps the ideal is multiply realizable,

given certain logically necessary tradeoffs between different values).[8] Decisions about how to reform the actual world, and which of the more ideal worlds to take as the goal, are a separate matter. On this view, the conceptualization of ideal worlds is useful simply in that it tells us what the political good is, and what justice is. Some of the less ideal worlds will be appropriate goals for change from the actual world. The second way of theorizing is to try to theorize *only* the less ideal worlds that are accessible from this one. So nothing fanciful or unrealistic, for example worlds where there is strict compliance with societal rules, should even be discussed. The problem with the second approach is that we are epistemically limited. We don't know for sure what is possible now, and it is difficult to predict what will be possible in the future given certain technological developments. So it will simply be more *difficult* to take the second approach, theorizing only up to the limits of what is realistically possible. If we are too cynical we will leave more just worlds unrealized; if we are too optimistic we will be guilty of idealizing in the sense criticized by Mills and others. So it just seems more prudent to take the first approach. That way we get the conceptual analysis for free, and we get a selection of increasingly better possible worlds from which we can choose a world, or aspects of a world, as a goal once sufficient investigation into our *ability* to transform our state of affairs into the desired state of affairs has been undertaken.

2.3 Defending ideal theory

In the last section I discussed several prominent criticisms of ideal theory, and suggested ways in which they might be answered. In this section I will draw on several recent defenses of ideal theory, in order to further strengthen the case for ideal theory. First I discuss the idea that ideal and non-ideal theory are not exclusive, including the idea due to Zofia Stemplowska that ideal theory might be necessary to political theory as a whole, the idea due to Adam Swift that ideal theory is permissible and perhaps necessary to certain kinds of work in certain disciplines, and the idea presented by both Richard Child and Federico Zuolo that ideal and non-ideal theory may play complementary roles within normative theory in general. Secondly I discuss the idea that ideal theory may simply be insulated from the kind of empirical criticisms that have generally been levied at it by non-ideal theorists. At this point I introduce Christian Barry and Laura Valentini's idea that principles are immune from empirical criticism,

8 E.g. people can't have perfect freedom without accepting limits to personal inviolability as a tradeoff.

and Gerald Cohen's idea that principles are fundamentally fact-insensitive.

2.3.1 The non-exclusivity of ideal and non-ideal theory

Adam Swift (2008) attempts to refute the criticisms of those who would have political theory give more practical guidance (i.e. function at the non-ideal level rather than the ideal) by arguing that there is an 'epistemological' and a 'practical' function in philosophy, the former being a question about truths, the latter being a question about actions. He argues that we do not hold other epistemological enterprises liable for failing to be practical:

> It is striking that we are less likely to criticize violinists, say, than political philosophers, for failing to provide justice-promoting guidance, as if being interested in identifying truths about justice meant that one was more rather than less culpable for failing to tell us how to bring it about (Swift, 2008, p. 367).

There is something jarring about the comparison between the political philosopher and the violinist. The political philosopher is *in the business* of talking about justice, after all. So let's take a clearer example: mathematics. Mathematics is a discipline that is both theoretical and practical. Research over the last thirty years in set theory is largely agreed to have no (foreseeable) practical application. Numerics, on the other hand, is used directly in encoding private information. The question is whether we should expect those dedicated to the 'epistemological' (theoretical) function of mathematics to contribute to its practical function. Should set theorists only be doing set theory if it can be shown to contribute to the practical functions of mathematics like data encoding? It seems that the answer should be 'no' – those are different parts of mathematics, and they are concerned with doing different things. Furthermore, it is impossible to know what the long-term implications of an apparently theoretical or impractical theory is – recursion theory started off as a theoretical project, and ended up giving us computers.

The application of this distinction to political philosophy is that philosophers of ideal justice should be under no pressure to contribute to the practical task of increasing the levels of justice in the world. The philosopher of justice is just like the set theorist in mathematics. The reason this might seem like such a strange claim is that normative subjects have a special subject matter that other subjects generally do not. People know what numbers are, and what it is for a scientific experiment to be controlled, and

roughly what it is to know something. But people have only vague ideas about what it is to be morally good, or what it is for a society to be fair, or just. In fact, justice is an extremely elusive concept. Given that we desire those things that are normative before we even know what they are, the 'theoretical' philosopher's job becomes important: tell us what justice is, so that we can bring it about! Swift is right that the philosopher's job is not necessarily *to bring it about*. But even so, her role is crucial in *it's being brought about*, because she is the one who must establish what, exactly, justice (and fairness, etc.) is. (This is to echo my answer to Sen's criticism of ideal theory in the previous section).

The 'epistemological' and 'practical' roles become blurred in normative theory in a way that they do not in the mathematics example, and that is a feature of the special subject matter with which political philosophy deals. (This is not to claim that if the political philosophers all died out there could be no justice in the world. But it is to claim that people in other disciplines, or people involved in the practical areas, e.g. the courts, the government, lobbying groups, political scientists, would themselves have to have a good idea of what justice was, or else we certainly would run the risk of things getting less rather than more just). To summarize, as in many other disciplines, normative political theory has a theoretical and a practical function. It is no complaint against theoretical philosophy that it is not practical; it was never meant to be practical. Theorists of justice are no more guilty for failing to make the world more just than set theorists are for failing to build safer codes for private data. Theorists of justice may even be less guilty, because on one understanding their job looks to be a necessary precursor of the practical job.

On Swift's view, ideal theory is immune from the criticism that it is impractical on the grounds that it is not the kind of thing that is meant to be practical. In her paper "What's Ideal About Ideal Theory?" (2008), Zofia Stemplowska argues in a similar vein that we should think of ideal and non-ideal theory as separate parts of complex political theories. She distinguishes ideal from non-ideal theory by defining non-ideal theory as that which issues recommendations that are both achievable and desirable, and ideal theory as that which does not. She argues that ideal theory is valuable because it can allow us to understand what our values require of us in the situations we find ourselves in, because it helps us to see through potentially distorting constraints, because it allows us to reflect on values which are complex and therefore non-transparent, because it can issue conditional requirements (*if* you want to bring about this kind of society *then* these conditions would need to hold) (Stemplowska, 2008, pp. 19-22). These characteristics do

not make ideal theory practical, but they make it valuable in a way that might be theoretically indispensable. Theory which fails to issue achievable and desirable recommendations, and does not instead make one of the contributions just listed, is, Stemplowska argues, *failed normative theory*, not, as others have argued, 'ideal theory' (Stemplowska, 2008).

Stemplowska's suggestion is that there's room for both ideal and non-ideal theory. This idea received further support at a recent conference on the relation between ideal and non-ideal theory in the United Kingdom. Pressure was put on the supposition that ideal and non-ideal theory are mutually exclusive, and that a theorist must choose which kind of theorizing to do, or that the discipline must agree to reject ideal theorizing altogether. Non-ideal theory is after all a *response* to ideal theory, with non-ideal theorists arguing that ideal theory involves idealizing assumptions which make it useless, irrelevant or inappropriate for the real world, and that we should for that reason simply stop doing it. But we might instead see the relation between the two kinds of theory as complementary.

For example, Richard Child has argued that there are two different approaches to justice, one axiological, the other dynamic. The axiological approach asks what justice is, while the dynamic approach asks what principles should govern the legitimate exercise of power (this is similar to Swift's distinction). He argues for a 'dual component' understanding of the two approaches, asserting that 'any coherent, plausible and complete conception of justice will have both an axiological component and a dynamic component' (Child, 2009). To some extent, Child's claim depends on our conception of a 'theory'. The logicians, for example, define theories as sets of sentences closed under logical consequence, a definition which imposes few constraints on how they are composed. More detailed definitions might include prescriptiveness, consistency, and explanatory power, might advocate the capturing of our basic intuitions, and so on. To the extent to which I understand the usage of 'theory' in political philosophy, as in 'a theory of justice', there is no inconsistency in supposing that there are both ideal and non-ideal components to theories' overall prescriptions.

Federico Zuolo makes a suggestion which is related but does not rely on a claim about how theories are composed. He argues instead that there are different *kinds* of theories, which have different aims. He separates 'intervention-oriented', 'justification-oriented' and 'reconciliation-oriented' theories, arguing that intervention-oriented theories propose models concerned with changing social reality; justification-oriented

models aim at establishing a standard of rightness (leaving aside the issue of how to implement that standard); and reconciliation-oriented theories aim to derive a standard of rightness from the world as it is (thus the question of implementing the standard does not arise, although the question of maximizing it might) (Zuolo, 2009). The idea is that how non-ideal a theory must be will depend on which of those kinds of theories it is. Justification-oriented theories may only need to be sensitive to the most general kinds of facts, while reconciliation-oriented theories will obviously need to take empirical reality seriously – how else to be sure that the standard of rightness thereby derived comes from the actual world? Zuolo proposes a model of different relevant empirical considerations, more of which will be relevant as we move from justification-oriented theory toward reconciliation-oriented theory.

The overarching suggestion here has been that there are other things, important things, for ideal theory to do than guide action. But one objection an opponent of ideal theory might raise to this suggestion that is that normative theory just *means* theory which is action-guiding. If that's true, then theory that doesn't directly guide action will fail to come out as a normative theory. Given that ideal theory doesn't usually guide action, all of ideal theory will come out as failed normative theory. But we ought not allow our opponents to simply stipulate their way to victory. Better would be to see what those engaged in normative theory construction understand their project to be, or what the most charitable and useful definition of a normative theory would be. We know that theories are often incredibly complex; just think of the forty years of research generated by Rawls' (1971). Stemplowska argues that a theory is 'a systematic account of our knowledge about a given dimension of reality ... that satisfies the criteria of what constitutes knowledge appropriate for that dimension' (Stemplowska, 2008, p. 5-6). She comments on the structure of theories that 'they have inputs, such as assumptions, outputs, such as final principles, and rules regulating the *derivation* of the output from the input' (Stemplowska, 2008, p. 6, her emphasis).

On this definition, it is perfectly consistent with being a normative theory that the final principles (the output) are not directly action-guiding. Some of the outputs might be what we would normally class as 'ideal' while others might be what we would normally class as 'non-ideal'. There is room in a given normative theory for it to function at both the ideal and the non-ideal level. Something that is not action-guiding in the immediate future might nonetheless have an effect on guiding action at some later time; or it might not ever have an action-guiding effect. Either way, we shouldn't close off the

debate about ideal and non-ideal theory by stipulating a definition of 'theory' that excludes the ideal in advance.

2.3.2 Ideal theory as fundamental or immune from empirical attack

Consider a broad moral principle like global egalitarianism, which holds that 'at a fundamental level, justice places limits on permissible global inequalities' (Barry & Valentini, 2008). Barry and Valentini argue that this kind of principle is immune from considerations about how the world actually is. We can think of ways of arranging the world such that the principle is realized to greater or lesser degree, and those hypothetical states of the world will be the object of considerations about implementation. But the principle itself will not be; the principle itself can never be something that is 'too unrealistic' or 'too impractical'. On this view abstract moral principles are such that they are forever insulated from empirically-based criticism. No matter how damaging a criticism about one way of realizing them, there is always the logical possibility that there will be another way.

This analysis is false for two reasons. The first is that surely even the most abstract moral principle should take *some* empirical constraints seriously. If a theory proposes a society in which every person earns more than the national average (to borrow again the example from Gilabert, ms., p. 11) then we should be able to reject it out of hand, not only as unable to be implemented but as impossible (one of the rare cases that we can rule-out, rather than just declare comparatively less feasible). Every way we can come up with of instantiating that principle in the world turns out to be impossible. Every way we can come up with of instantiating it in the world will be impossible *as a matter of necessity*, because the principle is logically contradictory.

The second reason it is false is that although there are lots of principles which are multiply-realizable, meaning that there are many ways the world could be such that they are realized, there are some which have more limited potential to be brought about. Consider the strictest possible formulation of global egalitarianism: 'everyone ought to have equal access to certain important advantages, thus avoiding relative deprivation' (Gilabert, ms., p. 2). Now consider conditions of scarcity, such that if we make it the case that everyone have equal access to certain important advantages, everyone will certainly die. If the world is such that scarcity is a persistent condition, and the principle in question is something strict like perfect equality of access for all, then there really is only

one way the principle can be realized, namely by awarding perfect equality of access to all. But under such conditions that is infeasible – it would be practically impossible to divide food up into small enough pieces to give everyone the same amount, and it would be irrational to let everyone die when there is a chance of saving some. If we know all of the ways a principle could be realized in the world, and we know that our actual non-ideal conditions would make all of those ways fail, then we know that the principle itself is unrealizable.

To restate the case, one reason for thinking that an analysis of principles in which they are immune from empirical criticism fails is that even the most abstract of principles should take *some* constraints seriously – logical possibility for one, but there are others. Another reason for thinking that such an analysis fails is that we sometimes know all the ways a principle can be instantiated in the world. Granted, we often do not; but when we do, the fact that all such ways are ruled out in practice is enough to show that the principle itself is ruled out. Barry and Valentini's 'immunity' defence of ideal theory is for those reasons unsatisfying.

More compelling, although still not entirely convincing, is Gerald Cohen's line. Political philosophers interested in the ideal theory debate have recently taken it to be describable in the terms proposed by Cohen (2003). The thought goes something like this: ideal theory abstracts away from the real world in a way that ignores the facts about how people and things are. So ideal theory is theory which is not sensitive to facts, and non-ideal theory is theory which is sensitive to facts. If we think of it like this, we can imagine a spectrum of fact-sensitivity, and place theories along it according to how seriously they take the empirical state of the world. Perhaps theories that ignore the real world entirely are fact-insensitive in the pernicious sense, because they legislate principles for people totally unlike us, while theories that take the real world absolutely for granted are fact-*sensitive* in the pernicious sense, because they do not legislate principles at all. So at the extremes, both ideal and non-ideal theory (if the mapping is legitimate) can fail; ideal theory can be too ideal, and non-ideal theory can be not ideal enough. If that thought is a reasonable one, then the fact-sensitivity debate is relevant to the discussion of ideal and non-ideal theory.

On Cohen's distinction, 'facts' are truths that might be taken to support principles, and 'principles' tell an individual what he or she ought to do. He argues that for any fact that supports a principle, there will always be some *further* principle that explains why the fact supports the principle, and that further principle will be ultimate.

He takes this argument to work against the popular assumption that all sound normative theories are 'fact-sensitive'. His claim is that ultimately, our principles must be *insensitive* to facts. Which is to say, ideal theory is unavoidable.

To get a sense of how his claim works, take the normative principle 'one should not steal unless in very serious need'. This principle might be supported empirically by facts about the consequences of stealing for society, or for the thief's reputation, and so on. Cohen's claim is that citing such facts to support the principle is far from the end of the matter. He thinks that when we ask *why* those facts support that principle, we'll give some further principle, which is not reliant on facts for support. So if we were to ask 'why do facts about the negative effects of stealing upon a person's reputation support the principle that they ought not steal unless in very serious need?', on Cohen's view the answer would be 'because we endorse the further normative principle that it is good to cultivate a good reputation'. Now in some cases, that further normative principle, at one remove from the initial pair of principle-plus-fact, will be the end of the matter. In other cases the chain will continue, but will ultimately, Cohen asserts, end in a principle. Continuing with the same example, one might ask why it is good to cultivate a good reputation, and the answer might come from appealing to facts about the benefits a good reputation bestows upon a person. But we can still ask why those facts support that principle, and in the end we might come to the normative principle that 'it is good for persons to be able to pursue their projects' (which a good reputation allows), which is something we just have to take as fundamental.

But at this point it becomes obvious that what is at stake for Cohen is not the same as what is at stake for the ideal theorist. Cohen's argument is that principles are *more fundamental* than facts. Both play a role, but the end of the chain, as it were, is a principle. But why should ideal theorists be concerned with fundamentality? The issue for this chapter is not which of ideal and non-ideal theory comes first, but whether there is or isn't some useful and valuable work to be done by ideal theory. On Cohen's story, the usefulness of fact-insensitive principles seems to be their explanatory relevance. The principles explain why the facts are relevant, and they let us distinguish between principles that are true because they are supported by certain kinds of facts, and principles that are asserted simpliciter.

In any case, we might try to identify principles simpliciter and principles supported by only the most general facts with ideal theory. Then we might still ask whether Cohen's argument succeeds, and thus whether ideal theory thus understood is

guaranteed a place in normative space. There are two related questions we might ask. The first relies on the distinction between unconditional (it ought to be the case that, if *p* then *q*) and conditional (if *p*, then it ought to be the case that *q*) principles. The question is 'do conditional principles always end in unconditional principles when we ask for an explanation as to why we hold them?' (For discussion on this issue, see Cohen, 2003; Cohen, 2008; and Pogge, 2008). The second concerns principles and their truthmakers. The question is 'what makes it true that principles, whether conditional or unconditional, apply?' With respect to the first question, the difference is only in whether a principle is true at all possible worlds, or true only at some.

If principles are unconditional then they are true everywhere: no matter where we are, it ought to be the case that if *p*, then *q*. But if principles are conditional then they are only true sometimes; if we are somewhere that *p* is true, then it ought to be the case that *q*. But nothing much hangs on which of these is true, practically speaking. At our world, *p* either will or will not be the case. If it is, then it ought to be that *q*. And this is true regardless of whether it ought that, had it not been the case that *p* (counterfactually), it still ought to be the case that *if p*, then *q*. Setting aside these scope worries, we should notice that almost all principles are conditional, maybe all. It's tempting to think of principles as asserting categorical requirements, such as 'you ought not harm other humans arbitrarily'. But as a matter of fact, these kinds of principles inevitably turn out to have built in caveats. What is really required is something like 'if you are a normally functioning human, then it ought to be the case that you do not harm other humans arbitrarily'. Perhaps the 'normally functioning human' clause is implied everywhere, so ends up being dropped in ordinary utterances. But that doesn't mean we should confuse the minimally fact-sensitive principle '(if you are a normally functioning human) it ought to be the case that you do not harm other humans arbitrarily' with the fact *insensitive* principle 'it ought to be the case that you do not harm other humans arbitrarily'. Obviously our moral principles don't apply to animals, or inanimate objects, or even aliens (David Miller makes a similar point in his (2008)). So it is not clear that they are ever completely fact-insensitive, and thus that principles are fundamental, even if only explanatorily so.

The second question looks superficially more promising, because if we *knew* what the truthmakers of principles were, then we could commit to meta-ethical fact-sensitivity. And that would tell us that there are at least some bad kinds of ideal theory – those that take *no facts at all* seriously. But questions about whether there are truthmakers

for moral truths, and what they might be, are part of a meta-ethical debate that has been running for an awfully long time with no real consensus. All this aside, we might notice that Cohen's argument does succeed, albeit on a technicality.[9] He agrees that principles are nearly always conditional, and builds this into the fundamental principles. If certain facts obtain, then the normative principles in question will be correct. We end up with a conditional: 'if fact *f* then principle *p*'. But conditionals themselves are logical entities, not 'facts': conditionals are fact-insensitive. So the principles that end up being fundamental on Cohen's story are fact-insensitive. Thus, the argument in favour of the fundamentality of fact-insensitivity is vindicated. But the fact *that* the principle (the ...'then *p*' part) is conditional upon certain facts obtaining (the 'if *f* ' part) makes the principles practically parasitic on the facts: the principles will be very different across different empirical contexts.

What does this show for the ideal theory debate? Non-ideal theorists argue that political philosophers should take facts about the actual world more seriously in building theories. The fact-sensitivity debate makes clear that principles can hardly do without facts. Even the most general and inclusive of principles will only apply to certain kinds of creatures, in certain kinds of contexts. Principles are almost always conditional, and what they are conditional on is who they apply to, and in what situation they are supposed to apply. So the interesting question is not whether principles are fact-sensitive, but how fact-sensitive principles are or ought to be. All theories will be at least minimally fact-sensitive. In Section 2.3.1 I introduced the idea that how constrained by the facts a theory should be depends on what kind of theory it is. In Section 2.4 I expand upon that idea.

In summary, we have considered the ideas that ideal theory is not vulnerable to attack on the grounds of being impractical because it was never supposed to be practical, that it may even be necessary to non-ideal theory, that ideal and non-ideal theory just play different theoretical roles (whether they are different parts of a political theory, or whether either one is sufficient for a political theory), and that a theory will have to be fact-sensitive to different degrees depending on what kind of theory it is. I think these suggestions go a long way toward illuminating the place of ideal and non-ideal theory within political philosophy. In the next section I want to present a two-stage view of theory construction, arguing to situate the importance of feasibility within it.

9 I am grateful to Jonathan Wolff for helpful discussion on this point.

2.4 The big picture: a two-stage view

We have surveyed the case against, and the case in favour of, ideal theory. I have argued that most of the criticism against ideal theory can be overcome. In this section, I will present a two-stage view that accounts for the place of ideal and non-ideal theory within political philosophy, and locates the feasibility project that will be the driving concern of this thesis within non-ideal theory. I will argue that there are two distinct levels at which facts about the actual world must be taken seriously.

Let's distinguish two stages of theory construction, from conception through to instantiation in the actual world. We'll call them Stage One and Stage Two. Facts about how the actual world is do enter the picture in Stage One, contrary to the beliefs of some contemporary ideal theorists, but only a restricted set of facts. These are broad empirical facts about what kinds of creatures we are. By 'we' I mean humankind. From this kind of radically minimal fact-sensitivity, we can derive general moral commitments. These commitments are likely to be pluralist, and might include such things as reducing the incidence of pain, increasing the possibilities for choice, and so on. Minimal fact sensitivity at Stage One is justified by any thought-experiment involving the moral commitments of creatures from distant planets.

It seems fairly clear that general moral commitments will differ between peoples of radically different constitutions, that a population who feel no pain will not find themselves committed to reducing pain ('reducing what?'), that a population who share a mind will not find themselves committed to increasing the possibilities for choice (Orson Scott Card's (1985) *Ender's Game* series has a nice exploration of these issues). Thomas Pogge makes a similar point against Gerald Cohen in his (2008). He comments that in order to be entirely fact-insensitive, Cohen's principles must apply in all possible worlds; but if that is so, it limits the number of fact-insensitive principles we can be sure of. For who are we to legislate our principles for creatures like us in all possible worlds? (Pogge, 2008; see also Miller, 2008). Of course, those who have no patience for far-fetched thought-experiments, will likely think this Stage One minimal fact-sensitivity is unnecessary, and that the scope of political theory is implicitly restricted to creatures like us. In that case, it is possible to take the general moral commitments for granted, and direct one's concern only at Stage Two.

In Stage Two, the requirement to take the facts about the how the actual world is seriously is much more stringent. There two things happen. In the first place, the general

moral commitments, which are subject to minimal fact-sensitivity, are translated into theories about what we are actually obliged to do. These will proliferate; there are many and varied ways in which to realize general moral commitments. In the second case, these theories, or normative recommendations, will be subject to stringent feasibility tests, which is to say, the theories at Stage Two will be maximally fact-sensitive. It must be shown that the prescriptions of a theory are realizable in the actual world before that theory can be adopted in practice. Although I have said nothing as of yet about how that process works - that will be the concern of subsequent chapters - I would suggest that the outputs of the process determine what is to be done. If there is only one theory whose recommendations pass the feasibility test, then that is the theory we ought to adopt in practice. If there are many theories, then we should choose the best among them. Notice that this picture stands in contrast to a view like Adam Swift's (2008). He argues that it is for the social scientists to determine what is feasible, and then for the political theorists to rank the feasible options, and to decide which of them (if any) to actually try to instantiate in the world. On my picture he misses a crucial step, which is that at Stage One it is for the political theorists to *figure out what to subject to the feasibility tests*. The philosophers are the dreamers; but the dreams must be made practical (or discarded if they can not be) before they can be realized.

Notice that the conclusion about which theories we should actually adopt in practice is conditional: *given* the general moral principles we have, and *given* the empirical context that makes realizing some of the theories more feasible than others, *these* are the theories we should adopt. There are many interesting questions to ask about this picture – in Stage One, what we can say about 'creatures like us' (and whether we can say anything that's both universal and useful) and meta-ethical questions about what the general moral commitments are, and whether (or how) we can access them; in Stage Two, how the stringent feasibility test works, i.e. what the relevant empirical facts are that the theories must be sensitive to, and along which parameters we are to establish 'best' when we must choose between competing theories which have all been shown to be realizable in the actual world. There are also interesting questions about whether this picture is top-down or bottom-up: do we need to conceive of a perfect world anywhere in Stage One to figure out what the theories in Stage Two should be? Or can we extrapolate from where we are, and what our general moral commitments are, in order to generate theories sufficient to run through the feasibility test at Stage Two? You might think we can sacrifice epistemic access to meta-ethical truth in favour of convergences

(or overlapping consensuses, in Rawlsian terms) in folk morality. If we can do things bottom-up, we have reason to deny at least one version of the ideal theory view, which is that ideal theory is somehow *necessary* to non-ideal theorizing, by giving us a goal, or a standard, or a way to measure progress. I have been inclined in this chapter to assume the top-down view, criticizing Sen's bottom-up view as unmotivated. But the matter is far from settled.

There are risks with a picture like this. One is that a general scepticism about the possibility of knowing whether a theory is realizable in the actual world might doom the entire project. If there's nothing we can say about whether the recommendations of one theory are more feasible than the recommendations of another, then Stage Two does no work at all, and there's very little point in even carving up the space into the two levels with a different stringency of fact-sensitivity at each. Assuming we are not beset by such a scepticism, another worry is that if we make the empirical facts which a theory must be sensitive to at Stage Two too stringent, then we risk capitulating to the mentality that what we have is the best we *can* have. We have to walk a fine line in telling the story of Stage Two fact-sensitivity, between demanding too much of people in a way that makes a theory unrealistic and impractical (because not sensitive to their actual capacities), and demanding too little of people in a way that makes a theory cynical and only weakly normative. It is the fine line demanded by Brennan and Pettit (2005); that we treat people neither as perfectly virtuously motivated, nor as terrible knaves.

Granting that this is a reasonable way to understand the picture of discussions about ideal and non-ideal theory, what are the implications for discussions about feasibility? The first is that 'feasibility' aligns itself most naturally with the maximal fact-sensitivity of Stage Two, not the minimal fact-sensitivity of Stage One. It's unlikely that we'll come across many political proposals that fail to be sensitive to the general facts about 'creatures like us' that are relevant to Stage One. The second is that if we're going to pick an area of the picture which needs work and where there's a good chance of making some progress, it's Stage Two. Stage One poses a familiar and well-worn picture; indeed, questions about the right moral principles are the questions moral philosophers have been trying to answer all along. And unfortunately, we are no closer to settling the meta-ethical truths than we were when we started (some would say we are much further away, given the burgeoning evolutionary evidence about the inculcation of moral norms and sentiments). So if Stage Two is reliant on Stage One, we'll likely have to take a pragmatic approach to settling the content of our general moral commitments, e.g. by

indexing them to the commitments the global folk have in common. Thus while we can agree that these commitments will likely be pluralist, we'll do better to stay out of that discussion entirely.[10]

But Stage Two is under-explored; there's a growing awareness that we want our normative theory to be more sensitive to empirical facts, especially where the theory is intended to be directly action-guiding, but the structure of that sensitivity has been for the most part not commented upon. This silence is not an artifact of there being nothing to say; indeed it seems that there is a lot to say and to be sorted out. Is feasibility a tool for ruling out, or for declaring more or less likely? If the former, what kind of impossibility is sufficient to rule a theory out as infeasible? If the latter, with what kind of probability do we establish likelihood, and how likely is likely enough to act on? What kinds of considerations are relevant to deciding whether the recommendations of a theory are likely – and what kind of view should we take upon human capacity for change? What if something is unlikely now, but we could do things to make it more likely in the future? And so on. It is to these interesting questions we will turn in the coming chapters.

2.5 Conclusion

I have argued that discussions about ideal and non-ideal theory can be seen as the wider issue inside which discussion about feasibility is nested. I proposed that we understand the debate about fact-sensitivity as involving two levels – Stage One, where general moral commitments combine with a minimal fact-sensitivity to form theories about how the actual world should be, and Stage Two, where those theories are subject to a stringent feasibility test, and where we choose the best among those theories that emerge as that which we should adopt in the actual world. I commented that there is more hope for making philosophical progress at Stage Two than at Stage One, given that issues in the latter have been the focus of so much past work, and issues in the former are under-theorized. Questions about how feasibility tests works are important in filling out the story of how general commitments become instantiated theories, and will

10 In a paper co-authored with Pablo Gilabert, I argue that there are actually three stages to theory construction: stage one, involving the formulation and defence of core principles, stage two, involving the design of institutional schemes implementing the principles from stage one, and stage three, involving strategies of political reform leading to the implementation of the institutional schemes from stage two (Gilabert & Lawford-Smith, manuscript). This section of our paper draws on and significantly revises the earlier discussion in (Gilabert, 2008, pp. 412-414).

also be much more important if it turns out that we can take a bottom-up approach to normative theorizing. Thus the task for the coming chapters will be to argue for a certain view of the Stage Two feasibility test (which we will from now on just discuss under the label of 'feasibility').

Chapter 3

Political Feasibility: Background Literature

3.1 Introduction

In this chapter I want to do two things. The first is to give an illustration of the way feasibility is used in political philosophy, to demonstrate the centrality of its role. To this end I shall discuss its role in work by Philip Pettit, Thomas Pogge, and Leif Wenar. The second is to notice that while there has been no extended discussion of the concept of feasibility, various aspects of it have been defended by a handful of authors, including Geoffrey Hawthorn, Pablo Gilabert, Allen Buchanan, Mark Jensen, Gerald Cohen, and Gillian Brock. I shall survey those accounts, discarding some, and taking elements of others through to Chapters 4 & 5, where I shall put forward a positive account of feasibility.

3.2 What role do claims about feasibility play in political philosophy?

In this section we will look at some of the ways that the concept of feasibility is used in political philosophy. This should suffice to demonstrate that the concept plays a central role, and that there is a lot to be gained in getting clear about what it means, and what it would take to establish the truth of a claim that some non-ideal theory is either feasible or infeasible.

3.2.1 Philip Pettit

In his (2002) Philip Pettit asks whether it is politically feasible to have a criminal justice system that is truly just. He notes that the current system of punishment doesn't seem to answer to any given rationale, and argues that that is because of the 'outrage dynamic' present in society. The dynamic comprises three stages: exposure, outrage, reaction. Some 'evil' will be exposed to the public, for example terrible conditions for workers in factories; the public will be outraged about the evil and demand that it be remedied; and democratically-elected representatives will be forced to act to address that evil, on pain of censure at the next election. Pettit's argument is that this dynamic forces an equilibrium toward harsher punishment in the criminal justice system, because

criminal acts are just the kind of thing to provoke outrage in a populace, and the media has an incentive to expose criminal acts, namely to sell more papers. His suggestion is that we should take criminal sentencing policy out of the hands of our elected representatives, so that sentencing is not subject to democratic pressure, and give it to an independent body, made up of e.g. crimonologists, lawyers, and members of the community.

What's important is not the detail of the argument but rather the way in which Pettit uses the notion of feasibility. He observes an empirical fact, namely that the criminal justice system conforms to no rationale, and proposes a hypothesis to explain that fact, namely the outrage dynamic present in democratic society. The title of the paper is 'Is Criminal Justice Politically Feasible?' Pettit's asking whether there's any way we can reform the criminal justice system to make it more just; whether there are changes *we* can make to *our* world (so not whether a more just system is merely possible). His argument is that such change may well be feasible, so long as we're prepared to take criminal sentencing out of the hands of the legislature. But he's skeptical:

> I identify one institutional arrangement that might make it politically feasible to shape the penal system by reasoned debate. I pin my hopes on the possibility of putting this sort of arrangement in place, for short of achieving it I am very pessimistic about the prospects for a decent system of criminal justice (Pettit, 2002, p. 2).

There are of course ways to criticize the view, namely to ask what the drivers of outrage in the community are, for this will have an effect on what the best solution to the problem will be. Is it, as Pettit argues, the fact that in a democracy our representatives are forced to respond to outrage? Or might it be instead that media sensationalism plays a role, or even the psychology of humans in groups, such that they respond particularly voyeuristically to violence, and encourage one another in excessive responses to crime? If either of those is the case, we might be compelled to ask whether reforming the media, perhaps by having greater state control and less private-sector media conglomerations, or educating the public, perhaps by teaching that moderate sentencing can produce better results overall, will be feasible alternatives to taking criminal sentencing out of the hands of the legislature. In Pettit's paper, it looks like 'feasibility' is about what options for change are available to us. The fact that he settles on the particular solution he does is not evidence that he thinks the other alternatives I mention here are less feasible or even infeasible (which would give us some idea of how his concept of feasibility worked); he has emphasized in his paper that the main driver of the outrage dynamic is the fact

that democratic representatives must respond strongly to public outrage, and that this will force an equilibrium of severe criminal penalties. So it's reasonable that he will focus on ameliorating the effects of that driver, rather than look to other possibilities. Feasibility is important because it helps us discover whether we can eventually have justice in criminal punishment, or not.

3.2.2 Thomas Pogge

Pogge (1992) argues for dispersing political authority over nested territorial units, as a means of introducing cosmopolitan institutional reforms. He gives feasibility a central role:

> We … consider the existing global institutional scheme unjust insofar as the pattern of human rights fulfilment it tends to produce is inferior to the pattern that its best feasible alternatives would tend to produce (Pogge, 1992, p. 54).

Pogge proposes a standard by which we are able to figure out whether or not our current global institutional scheme is just. The standard is *the justness of the best feasible alternative*. The process by which we would assess whether a current global institutional scheme met that standard would be to (a) put on the table all of the non-actual, but feasible, alternatives to that scheme, (b) assess each of those alternatives in terms of how well they fulfill human rights claims, and (c) rank each of the alternatives, including the actual-world scheme, according to (b). If the actual global institutional scheme comes out as inferior to any of the alternatives in (a), then it is for that reason unjust. If it comes out as superior to those alternatives, then it is presumably just.

Most of what is required in using this standard to assess the justness of a global institutional scheme is empirical considerations. We need to know what the alternatives are, and, given some pre-defined notion of human rights, we need to know to what extent each of the alternatives, including the actual scheme, fulfill them. You will notice that a lot hangs on what counts as a 'feasible alternative'. If philosophers have different ideas about what it is to count as a feasible alternative, then their conclusions about whether the current global institutional scheme is just or not will diverge accordingly. Thus it seems to be fairly important that our notion of feasibility is well worked out.

3.2.3 Leif Wenar

In his (2006) discussion of Rawls's *Law of Peoples*, Leif Wenar argues that cosmopolitanism, on the version that most contemporary writers prefer (a globalized version of the domestic original position familiar from Rawls's *Theory of Justice*, in contrast to the statist form of cosmopolitanism that Rawls himself actually defended), is *impossible*. This is a bold claim. Wenar argues that a global state with a monopoly on power and the use of force is either impossible or undesirable, and that without such a global state, we are left with territories and territorial defence, which is to say, not the form of cosmopolitanism that cosmopolitans prefer.

There are some strange things going on in this argument. The first is the claim that cosmopolitanism in the preferred form is impossible, which is supported only by the step in the argument which claims that it is *either* impossible *or* undesirable. But the fact that a theory proposes a state of affairs which is undesirable has no bearing on whether or not that theory can be implemented in the world. Plenty of undesirable states of affairs have been brought about in recent history. For the preferred form of cosmopolitanism to be impossible, it must be demonstrated that there is no way of bringing it about. But surely there *is* a way of bringing it about; it's just that we don't want it brought about on those terms. For example, a powerful state like Russia or the United States might use violent coercion to claim world power, under threat of massive global nuclear attack. This is not a way we'd like the world to be, but it's not clear that there's anything ruling it out. The idea behind Wenar's claim that cosmopolitanism is impossible is that there are certain precursors to it which we have little hope of meeting. These are, most importantly, peace and security. While there is war, and while people are unsafe in their own territories, there can be no global agreement of the kind secured in the domestic original position.[11]

The details of Wenar's claim are not as important here as the structure of his argument. He shows in the paper that there are two kinds of cosmopolitanism, the one preferred by most contemporary writers which simply makes the domestic original position with its focus on individuals an international one, and Rawls's own statist version, which focuses on *peoples*. Then he argues that the kind of world imagined by the former is impossible, because the precursors of a world like that look nowhere imminent, and because to sustain it we'd need a global state with a monopoly on power,

11 I am grateful to Leif Wenar for clarification on this point.

which would be undesirable, and maybe even unachievable. We can disagree with these claims (perhaps a global state with a monopoly of power is *not* necessary; perhaps power can be distributed federally without that forcing us into Rawls's statist version of cosmopolitanism), but the key thing to notice is the crucial role that feasibility is playing in this argument. A popular version of cosmopolitanism is being rejected on the grounds that it is infeasible – even more strongly, on the grounds that it is impossible. The success of the argument depends on whether this claim is true. But we cannot evaluate the truth of the claim if we have no idea what it would take to establish it.

3.2.4 A central role

Politicians and political philosophers are forever painting pictures of better worlds and encouraging us to commit resources to moving toward them, whether recommending small improvements such as better recycling systems, or revolutionary changes such as the movement from free-market capitalist societies to perfectionist ones (e.g. Raz, 1986), or to socialist ones (see discussion in Cohen, 2009; Gilabert, forthcoming). For any such proposal, we are in a better position to comment upon it, and ultimately reject it as a waste of resources, or better suited to the science-fiction bookshelf, if we have a good idea of *what it takes* for a proposed outcome to be a feasible one.

Without an idea of what it takes, politicians, government officials, political scientists and philosophers risk talking past one another. More frustratingly, they risk talking past each other because of differing ungrounded assumptions about the capacities of human persons. It would not be surprising to find, in the historical political literature, a disagreement between key figures underwritten by their conflicting views about 'human nature', where one is optimistic about human potential, and the other pessimistic or cynical. You might indeed think that some of that kind of disagreement is going on between John Rawls (1971) and Robert Nozick (1974) in their famous exchange. Disagreement about the nature and capacities of persons won't be entirely eradicated by providing a notion of feasibility, but it may be at least partially ameliorated.

David Held asserts the importance of the concept, although not by that name:

> Today, any attempt to set out a position of what could be called "embedded utopianism" must begin from where we are – the existing pattern of political relations and processes – and from an

analysis of what might be: desirable political forms and principles (Held, 1992, p. 345-6).[12]

He stresses the importance of starting from where we are now, and analyzing 'what might be'. Such an analysis requires taking real-world constraints seriously, and that is to employ a notion of feasibility.

> …the question of feasibility cannot simply be set up in opposition to the question of political ambition. For what is ambitious today might be feasible tomorrow. Who anticipated the remarkable changes of 1989-90 in Eastern Europe? Who foresaw the fall of communism in the Soviet Union? …The question of political feasibility is of the utmost significance (Held, 1992).

For the remainder of the chapter I will go on to look at authors who have defended aspects of the concept of feasibility, in order to establish what the necessary elements are. The project of the next couple of chapters is primarily one of explication, in the Carnapian sense (Carnap, 1950, p. 3).[13] I want to defend a version of feasibility that will be *most useful* to political philosophers engaged in debates over whether one another's theories are appropriately 'non-ideal', 'realistic', 'non-utopian', or 'practical'. That means I might end up defending a concept that departs from the way the word 'feasibility' is used in ordinary language, and I might end up defending a concept that departs from what the concept "really" means, on a Platonic story about the nature of concepts. My aim is only to give some criterion by which disputes over how realistic some theory is can be more or less settled.

3.3 Feasibility and its cousins, in the literature thus far

In this section I want to survey a couple of attempts in the political philosophy literature to specify aspects of an explicit notion of feasibility (or a closely related concept). There are few such attempts. I will concentrate on Geoffrey Hawthorn's (1991), Allen Buchanan's (2004), Pablo Gilabert's (2009), Mark Jensen's (2009), Gerald Cohen's (2009), and Gillian Brock's (2009) accounts, because I think I can take something from each of them, and move forward in Chapter 4 to build out of those ideas a positive account of my own. While I think there is something promising about each of them, I think none specify a full account of political feasibility (although Cohen gets close).

12 Held refers to the Real Utopias Project of Erik Olin Wright. See Wright, 2010.

13 "The task of explication consists in transforming a given more or less inexact concept into an exact one or, rather, in replacing the first by the second. We call the given concept (or the term used for it) the explicandum, and the exact concept proposed to take the place of the first (or the term proposed for it) the explicatum" (Carnap, 1950, p. 3).

3.3.1 Hawthorn's *Plausible Worlds*

Hawthorn (1991) is concerned with counterfactual *histories* rather than counterfactual (and actual) *futures*.[14] Nonetheless, his discussion is instructive. He gives three main conditions necessary to our formulating plausible counterfactual histories.

The first is that counterfactual histories should 'start from a world as it otherwise was' (Hawthorn, 1991, p. 158). That is to say, we must hold fixed *everything* about the world up until the time of the event we wish to permute in order to derive counterfactual results. He states explicitly that such possibilities 'should not require us to unwind the past' (Hawthorn, 1991, p. 158).

The second condition is that the consequences or implications we take to follow from the permuted events 'should initially fit with the other undisturbed runnings-on in that world' (Hawthorn, 1991, p. 158). Presumably this means that we ought not need to change things in China when we permute small details in Australia; but this condition becomes increasingly difficult to defend in an age of easy global communication, trade, mass media, and so on. So it is not at all clear what it would be to 'fit' with the other undisturbed running-on in the world, unless this merely means that the implications of the permuted event are not to be in blatant contradiction with events happening elsewhere. But even that is problematic.

The third condition Hawthorn gives is that neither the changed events, nor their implications, should be 'fantastic' (Hawthorn, 1991, p. 158). The idea here is that given the momentum and inevitability of certain events, it would be altogether too fantastical to suppose that they could have been otherwise. This last specification functions as a constraint upon the logical space of *candidates* for permutation. For example, Karl Marx had such an impact upon people's thinking, and Russia's political leaders were so ideologically driven, that it was *bound to be the case* that Russia ignored Marx's criterion for the kind of country 'ready' to instantiate communist ideals, and went ahead with its disastrous attempt. Given certain facts about people, dispositions, temperaments and environments, some events are inevitable, or almost so. This third condition is supposed to rule out events departing from the inevitable from being candidates for counterfactual histories (futures). Even if we take the third condition seriously and disallow all 'fantastic' alternatives, there is still a question of how relevant alternatives are to be generated. Hawthorn gives three possibilities.

14 For an interesting book which develops counterfactual histories, see Tetlock et al., 2006.

The first is to count as an alternative whichever outcome a given set of facts *could have* yielded. So if a given event *A* could have yielded outcomes *x*, *y*, and *z*, where *y* is the actual outcome it yielded, then we ought to take *x* and *z* as the relevant alternatives. Another is to take event *A* and its actual outcome *y*, and consider states of affairs close to *y*. So if *y* is close to *y'* and *y''*, then we ought to take *y'* and *y''* as the relevant alternatives. One question Hawthorn considers is whether candidate alternatives should be all those that exist in the logical space of a given event, or only those actually considered possible by agents at the time of an event, i.e. the ways the world could have gone which were 'forseeable'. This is an interesting consideration but one where the answer seems fairly obvious: unless we want to be limited in our discussions to all the vagaries of actual agents and what they may or may not have considered a forseeable implication of some event or series of events (consider Hitler's secretary Traudl Junge, who claimed in a documentary before her death that she had not been aware of the 'extent' of events in Nazi Germany (Hirschbiegel, 2004)), then we ought to take logical space to be the answer. The last possibility Hawthorn gives is to leave it to some theory to generate relevant alternatives, although what that theory would be is almost exactly what is at issue in this thesis.

Later in his work, Hawthorn comments that at least in history, we can know a great deal about the dispositions and abilities, attitudes and prejudices, of individual agents, and that these considerations not only make it the case that many alternatives are not possible for them, but also make it the case that many alternatives are not possible for us to ascribe to them (Hawthorn, 1991, p. 166). The better informed we are about agents, the better we are able to get a handle on what is possible for them, and what is not. One problem (there are many) with transcribing these conditions for counterfactual histories onto counterfactual futures is that while we can know a great deal in many cases about historical figures, we can know almost *nothing at all* about the dispositions, attitudes, abilities and prejudices of future persons. What we can know is likely limited to very general attributes which all individuals are likely to have in virtue of the kind of future we cast them into, for example we can reasonably expect our descendants to be resentful of their ancestors for the lifestyles which triggered global climate change, but whether such general information is of much use remains to be seen.

The problems with Hawthorn's three conditions for figuring out plausible histories (feasible futures) are even greater than that. For one thing, although his is at least an *attempt* at rigidifying the methodology of counterfactual histories, there is very

little content to the three conditions he provides. If we already knew what was to count as 'fantastic' then we would not be in the business of looking for a structured approach to feasibility. A statement to the effect that 'what is feasible is just that which is not fantastic', although we admittedly have *some* intuitive grasp of its meaning, is more or less unhelpful.

Furthermore, the idea of holding fixed everything about the world up until the time of the event we wish to permute is problematic (David Lewis discusses this problem and introduces the notion of 'miracles' to allow the transition to the counterfactual future, see Lewis, 1979). Sudden changes in events are highly implausible if disjointed from the intentions of the agents who brought them about, or from environmental conditions. With respect to counterfactual histories, it has to be reasonable to suppose that things could have been otherwise – this means tracing a causal history back to a point where it is realistic that things turned out differently, and this is no easy task. With respect to counterfactual futures, we have to make assumptions about how the world is that might turn out to be false, e.g. that a populace is likely to be motivated by a moral campaign arguing for less cars on the road. In some sense, our epistemic ignorance about our social situation prevents us from perfectly fulfilling Hawthorn's first condition.

And as already mentioned, it is hard to make sense of the idea that the consequences of our permutations (i.e. considering possible alternative futures) should initially fit with undisturbed states of affairs elsewhere in the world. It seems that this condition runs the risk of being rather trivial, 'do not change things that would not be changed by your permutation'. Epistemic ignorance also enters the picture in that it is simply hard to know how widespread the ramifications of certain actions or events will turn out to be.

A strength of Hawthorn's account is his discussion of the inevitability of certain events. This is not to endorse any kind of metaphysical determinism, but rather to say that we should not ignore the momentum of certain social movements when we consider possible futures. If it is clear from the history of the last sixty years that the human rights movement is going from strength to strength, then we should ignore proposed futures in which human rights are suddenly no longer a concern. If we know a lot about the character, environment, and general dispositions of certain political figures, then we shouldn't posit futures that would require wildly out-of-character actions from them as a means of getting there. And if we know we are in a world still reeling from the failures of certain kinds of political 'experiments', e.g. communism in Soviet Russia,

racial purity in Nazi Germany, or racial assimilation in colonial Australia, then we shouldn't posit futures in which these 'experiments' have succeeded, because we cannot expect people to simply forget political history. (This is part of a point made by Daniel Dennett (1995) about biological possibility, namely that certain developmental paths that might have been open at one point in time are lost to us forever when the environmental conditions and hosts change or go extinct. At time t_1 a lot is possible that is not possible at times t_2, t_3, ...t_x, and that is something that any account of feasibility should capture.

3.3.2 Jensen's 'The Limits of Practical Possibility'

Mark Jensen (2009) argues that the crucial condition of the four necessary and sufficient to political feasibility (which he calls 'practical possibility') is 'natural human ability'. The other conditions he argues for – logical consistency, non-violation of laws of nature, and fixed history of the world – are identified rather easily as being met (or not) under some given proposal (more on these in Chapter 4). Natural human ability, he argues, 'garners the most attention' (Jensen, 2009, p. 8), and is therefore in need of the most discussion.

Jensen proceeds in his paper to separate natural human ability into three categories: synchronic, direct diachronic, and indirect diachronic. A person has a synchronic ability if she can perform an action *now*, a direct diachronic ability if she can perform an action now or later, and an *in*direct diachronic ability if she can perform an action later, provided that she perform another action first (an action which will enable her to perform the later action) (Jensen, 2009, p. 14; foreshadowed in Talja, 1985, p. 238). To give an example of these, I have a synchronic ability to work on this chapter, a direct diachronic ability to resume working on this chapter as soon as I restart my computer in the event of its crashing, and an indirect diachronic ability to rewrite this chapter in German, so long as I take some classes to improve my German before I do so.

Jensen argues that work on practical possibility is most concerned with the latter, namely, indirect diachronic abilities. This is probably because he thinks synchronic and direct diachronic abilities are transparent to both their author and others, although I suspect sometimes we are hindered in reaching conclusions about what is possible precisely because we are ignorant of the general abilities people have. Indirect diachronic abilities are obviously more difficult to predict: even if we are very good at knowing about the other kinds of abilities, the future is epistemically opaque when it

comes to these. How can we know what a person can do, if they choose to put themselves in a position to do it? Jensen argues furthermore that 'insofar as group actions mimic the actions of individual agents, we can assign our three kinds of abilities to groups' (Jensen, 2009, p. 16). That complicates even further the issue of indirect diachronic abilities – now we have to figure out not only what things individuals can do, should they put themselves in a position to do them, but also what things groups of individuals can do, should each of their members put themselves into a position to do them. The chains of cause and effect are far from straightforward.[15]

What's nice about this account is that it makes some useful distinctions for how we think about 'ability'. It points out that for something to be feasible it doesn't have to be the case that it is feasible *now*, it just has to be the case that it is feasible *at some point*. The introduction of indirect diachronic abilities makes precise the idea that we can do something now to make it the case that we can do something else later on; and so on *ad infinitum*. What's less satisfying about the account is that although we're now better at classifying *kinds* of abilities, we're no closer to knowing what kinds of abilities people actually have. The incredibly difficult issue in all of this is knowing what to say about proposals that require a huge departure from the way the world is now. We might be in a good epistemic position to agree that the early parts of the departure are possible, and in absolutely no position to comment on the later parts.

But we can resolve that criticism in favour of Jensen. There is little fixed content to be had that is both general enough and useful. He attempts to provide a framework for thinking about abilities as they matter for practical possibility. In the same way, I will attempt to provide a framework that could be used in deciding whether a proposed political outcome can be achieved from where we are. In Chapter 4, I will begin to develop that framework.

15 This also creates an interesting problem for normative theory, insofar as 'ought implies can' should be read as 'ought implies feasible', where feasible means something like 'can, conditional upon trying' (following e.g. Brennan & Southwood, 2007; on which more in Chapters 4 & 5) and we take Jensen's disambiguation of abilities to structure the 'can' concept. For if a person should do not only what they can do but what they could put themselves in a position to do, then moral duties proliferate, as does blame for not having done what one could have. To give an example, perhaps I am morally blameworthy for the fact that I cannot save the man on the street having a heart attack, because I could have chosen to go to medical school instead of doing philosophy, and I could have taken a first-aid course last week instead of going to school. But just think of how many such cases will be true if all feasible alternatives are allowed...

3.3.3 Cohen's *Why Not Socialism?*[16]

In *Why Not Socialism?* Gerald Cohen explores two questions: whether socialism is desirable, and whether it is feasible (Cohen, 2009). In trying to answer the question of whether it is feasible, Cohen makes three useful distinctions. The first distinguishes principles from their implementation (Christian Barry and Laura Valentini make a similar distinction, see Barry & Valentini, 2008). That is to say, it is not principles themselves that can be feasible or not, but rather their instantiation in the world given various constraints. There is something to the intuition that principles are 'immune' from feasibility constraints, but that distinction seems rather flimsy when we consider the logic of the matter. If a principle *p* can only be instantiated in the world in three different ways, *a*, *b*, and *c*, and we know that all of *a*, *b*, and *c* are infeasible, then we know that instantiating the principle is infeasible. It might not be good linguistic practice to say that the *principle* is infeasible, rather than being careful to point out that *implementation* of the principle is infeasible, but they amount to much the same thing in practice.

The second distinction Cohen makes is to separate feasibility from desirability. A state of affairs might be desirable without being feasible, and it might be feasible without being desirable (more in Chapter 5). The final distinction is between accessibility and stability. Both are important. There must be a route by which we can get to certain states of affairs under consideration, and the states of affairs themselves must be sustainable (Cohen, 2009, pp. 56-57; Cohen, 2001). He seems to think that the serious issues are limits to technology, and limits to human nature. He concentrates for the most part not on whether we can change people, but on whether we can change institutions in a way that they they can change people, or handle people unchanged. In the end he is agnostic about whether socialist principles can be fully implemented. He says that we do not know them to be feasible, but neither do we know them to be infeasible (Cohen, 2009, pp. 75-76; see also discussion in Gilabert, forthcoming).

Because we will largely set aside questions about desirability here, Cohen's first distinction, between principles and their implementation, is not particularly necessary for our discussion. His claim that feasibility and desirability can be separated in part justifies concentrating on one without the other. The most important thing is his claim that accessibility and stability are important. I will take these through to Chapter 4 as

16 I am grateful to Pablo Gilabert for bringing this work to my attention. The discussion in this section is informed by (Gilabert, *forthcoming*).

crucial ingredients of a theory of feasibility.

3.3.4 Buchanan's *Justice, Legitimacy and Self-Determination*

Allen Buchanan seems inclined to place feasibility constraints upon *ideal* theory rather than simply, as I argued in Chapter 2 we should do, leaving those as requiring only minimal constraints and focusing instead on the more stringent constraints that should exercise a limit upon non-ideal theory (Buchanan, 2004). He argues that ideal theory must be feasible, accessible, and morally accessible. He means something different by these terms than I will end up meaning. An ideal theory is feasible if 'the effective implementation of its principles is compatible with human psychology, human capacities generally, the laws of nature, and the natural resources available to human beings'. He thinks any theory that fails to meet this condition is of 'no practical import'. An ideal theory is accessible if there is a 'practicable route from where we are now to at least a reasonable approximation of the state of affairs that satisfies its principles'. Finally, it is morally accessible if the transition demonstrated in the accessibility condition does not involve 'unacceptable moral costs' (Buchanan, 2004, p. 61). These conditions Buchanan takes to be constraints upon *good theory*, not upon a concept of feasibility. His account of feasibility is only the compatibility with various facts about human capacities mentioned already. But it seems to be that we should combine his feasibility and accessibility to get an optimal condition (more on this in Chapter 4, and more on why we should resist a moral accessibility condition in Chapter 5).

3.3.5 Gilabert's 'The Feasibility of Basic Socioeconomic Human Rights'

Pablo Gilabert (2009) asks about the feasibility of basic socioeconomic human rights, including rights to food, clothing, housing, basic medical care, and basic education. He takes it that these are the conditions of a minimally decent life. Such human rights are uncontroversially desirable, but are not obviously feasible. The problem is that rights imply obligation or duties in other people, but people cannot have obligations or duties to do what they cannot do, on the standard understanding that (action-guiding) oughts imply 'can'. Gilabert discusses Maurice Cranston's (2001) argument that basic socioeconomic rights are infeasible on the grounds that they require

positive action on the part of others, rather than the 'refraining from' involved in e.g. civil and political rights. To provide civil and political rights the government must restrain itself from interfering with certain of citizens' choices; but to provide socioeconomic rights the government must *do* something, something that will often be complicated and costly. To reject this distinction and the alleged infeasibility of socioeconomic rights, Gilabert defends a notion of political feasibility.

He thinks feasibility comes in different types, domains, and degrees. There are two types: minimal, and expansive (these correspond to Brennan and Pettit's 'hard' and 'soft' constraints (Brennan & Pettit, 2005). For something to be minimally feasible, it must be logically, physically, and biologically possible. For it to be expansively feasible, it must be economically, politically, and culturally possible. He notices that we must be careful about how seriously we take the latter, because it is not impossible to effect social change that overturns economic, political and cultural norms and practices. To that end it is useful to separate strict impossibility from mere improbability. Socioeconomic human rights might be infeasible because they violate minimal feasibility constraints, which renders them impossible, or because they violate expansive feasibility constraints, which renders them improbable. The upshot is the same, but the reasons for why it is true are different.

As with Jensen's distinction between static and dynamic constraints, Gilabert agrees that there is an important difference between constraints that are fixed, and constraints that are malleable (Gilabert, 2009, Section III). And as with Cohen, Gilabert distinguishes the domain of feasibility as including considerations about stability and accessibility, namely, whether a desired outcome will be stable, and whether there is a way we can bring it about. Finally, with respect to degrees of feasibility, Gilabert argues that some outcome can be completely infeasible or completely feasibile (e.g. when it either clearly violates minimal feasibility constraints, or when it clearly satisfies both minimal and expansive feasibility constraints, respectively), but also that it can be feasible to degrees in between (e.g. when it meets expansive constraints to some degree or other). He notices that circumstances change, so that what makes it infeasible for a person to act now might not make it impossible for a person to act at some later time.

Gilabert seems to want an assessment of feasibility to be a matter of people's duties, e.g. to inform a judgement that it would be feasible for X to claim a right to Z, or for Z to fulfill his obligations in providing X with Y. This is slightly different in scope from what I want to do in this chapter, which is to give an analysis of feasibility for states

of affairs, or outcomes in the world, rather than for individual actions (although of course there is a state of the world in which some individual has undertaken some action). In the next subsection I will discuss Cohen's account of feasibility going on in Chapter 4 to pick up the elements introduced so far and put together a comprehensive notion of political feasibility.

3.3.6 Brock's *Global Justice*[17]

Gillian Brock (2009) aims to provide 'workable' claims about the realization of global justice, addressing 'concerns about implementation', and allowing us to move 'from theory to feasible public policy' (Brock, 2009, p. *vii* & p. 4). She addresses two kinds of sceptic about the possibility of cosmopolitan global justice. The first kind of sceptic claims that we cannot do what the cosmopolitan claims we ought to do; the second kind of sceptic claims that we should not do what the cosmopolitan claims we ought to do. The second claim rests on the idea that doing what the cosmopolitan claims we ought to do would interfere with nationalism, and other goods like authentic democracy (Brock, 2009, Ch. 1). Let's focus on the descriptive claim that we cannot do what the cosmopolitan claims we ought to do, which is where feasibility is important.

One might expect that in a book addressing sceptics about the feasibility of cosmopolitan global justice there would be some kind of criteria given for what a victory, or a defeat, might consist in. In other words, one might expect at least a sketch (and at most an explicit account) of the conditions under which a cosmopolitan proposal is to count as feasible. But Brock resists an explicit account, instead choosing (we may assume) to rely on a commonsense or pre-theoretical notion of feasibility. I'm not sure that there even *is* a commonsense account, but I think we can figure out what concept Brock is assuming by looking at how her analysis works.

As far as I can see, Brock uses two main argumentative strategies. The first is something like *a fortiori*, arguing that certain things are feasible because they are actual, and the second is something like argument by generalization, arguing that certain things are feasible because they have been realized in part. In Ch. 8 ('Immigration') Brock discusses the fact that a high proportion of migrant workers send money back to their families in their home countries. These remittances have both positive and negative

17 Discussion in this section draws upon my review of Brock's *Global Justice* (2009). See (Lawford-Smith, forthcoming).

effects on the receiver countries. One effect which might be considered negative is that the remittance money is generally spent on daily consumables, rather than being used on public goods like health care, education, roads, and sanitation, a lack of which is the structural source of developing country poverty. As a solution to this problem, which is one obstacle to alleviating global poverty, Brock suggests that the countries employing migrant workers might compulsorily deduct a percentage of the workers' earnings to send back to their home country, and even better, the home country might match these funds 1:1 and use the raised money to provide public goods. Brock's evidence for thinking this kind of tax on remittances is feasible is that it is *actual*: 'as happens already in many cases of Filippino, Chinese, and Korean workers' (Brock, 2009, p. 207).

In Ch. 5 ('Global Poverty, Taxation, and Global Justice') Brock concentrates on global taxes that might be used to fund the relief of developing nation poverty. Much of the argumentation in this chapter, too, follows along the lines 'possible-because-actual'. A global tax is feasible, she assumes, because we have already partial success in implementing one, e.g. air-ticket taxes (where a small fee is added to the sale of aeroplane tickets, and which goes towards providing pharmaceuticals and malaria treatments in developing countries). Initially thirteen governments agreed to introduce air-ticket taxes, and now thirty-eight governments have implemented it (Brock, 2009, p. 133-134). Or for another example, we have had success in implementing a tax on deep seabed mining, implemented via a United Nations' convention in the 1980s and now signed by 158 countries (Brock, 2009, p. 131). Brock also comments that two popular proposals for a global tax, namely a currency transaction tax and a carbon tax, 'have achieved a small measure of implementation success' (Brock, 2009, p. 132). The carbon tax has been enacted in Sweden, Finland, Germany, the Netherlands and Norway; the currency transaction tax has attracted conditional commitment from Canada, Belgium and France (these countries have promised to enact the tax if there is support from the international community).

In Ch. 8, Brock discusses immigration and its relation to global justice. She argues that while it is unclear which of the argument for open borders based on a behind the veil preference for large-scale freedom of movement, and the argument for closed borders to protect cultural community, would prevail in ideal theory, it is rather clearer that in non-ideal theory, immigration poses a serious threat to those who are left behind (Brock, 2009, p. 191). She uses recruitment by developed countries of healthcare workers from developing countries as an example, discussing for example the dire

shortage of healthcare workers in sub-Saharan Africa, and the effect the shortage has on the population there, which is a direct result of such recruitment. Brock argues that a 'comprehensive solution' to this kind of problem involves the following:

> (1) an international code that specifies uniform standards for both private and public sectors, and that applies to all countries in similar circumstances; (2) an international agency that oversees activities, brokers compensation, can punish violators (perhaps by levying meaningful fines), and so forth; (3) each country's aiming at and achieving self-sufficiency with respect to human resources in health care; and perhaps (4) addressing the seemingly insatiable demand for healthcare in developed countries (p. 202).

The assumption is that these components of a comprehensive solution are individually and jointly feasible, and Brock suggests that we have made progress towards justice in healthcare worker recruitment by pointing to the fact that 'there is already at least one version of an international code that could do the job outlined in (1) and several proposals concerning (2)' (Brock, 2009, p. 203, ft. 41). She considers the World Health Organization, alone or in conjunction with another international organization, to be a good candidate for the job of (2).

Or for another example, in Ch. 7 ('Humanitarian Intervention'), Brock argues that there is progress toward the goal of having sovereign nations intervene with e.g. corrupt or human rights- abusing nations in order to restore justice to their citizens. This progress consists in the International Committee on Intervention and State Sovereignty (ICISS) having come up with clear guidelines for when an intervention is acceptable Brock, 2009, pp. 181-184). Again, this is a partial progress. If clear guidelines upon acceptable humanitarian intervention are the first step in creating a just world order in which states intervene upon other states for humanitarian reasons, then we have reason to think we have made progress toward that goal.

The argument extrapolating from part to whole and the argument moving from actual to feasible seem to amount to roughly the same thing, namely the conclusion that progress is good evidence of eventual success. But that conclusion is problematic, because it might just as well turn out that the features that make the first steps toward some goal feasible are exactly the features that make achieving the goal infeasible. An internationally-implemented luxury goods tax, or currency exchange tax, for example, are on a different scale and with a different scope to a thirty-eight government strong air-ticket tax. The fact that the carbon tax has success only among countries in the European Union (EU) might be one feature that makes it more likely to succeed within the EU and less likely to succeed internationally, because the EU has a central governing

body in a way that the wider world does not.

Likewise it is just as conceivable that coming up with guidelines on acceptable humanitarian intervention should count as progress toward the goal of actual acceptable interventions on humanitarian grounds as it is that it should end up illustrating more clearly that no real-world circumstances are likely to occur such that an intervention would count as acceptable. Just consider a parallel case: the fact that we can achieve relatively high levels of compliance with the state's laws in a modern democracy shouldn't be taken as progress towards realizing the goal of full compliance with the state's laws, nor as evidence that full compliance is something we can realistically hope for. There are circumstances in which partial success or progress can be taken as a sign that a goal is feasible, and circumstances in which it would be crazy to draw that conclusion.

Bill Gates has argued recently that this kind of 'progress' reasoning is detrimental in efforts to reverse the effects of climate change. His idea is that the goal of a 30% reduction in carbon emissions by 2020 is taken to be 'progress' toward the goal of an 80% reduction in carbon emissions by 2050, but that in fact the 2020 goal is realizable by small improvements while the 2050 goal requires radical new innovation. We do not get closer to the 2050 goal by reaching the 2020 goal, because none of the things needed to 'set up' realizing the 2050 goal will have been done. So in fact, he thinks, concentrating on the 2020 goal and the small improvements in efficiency it requires are detrimental rather than 'one step along the way' to the main 2050 goal. If our goal is an 80% reduction in carbon emissions, then 'progress' shouldn't always be taken as linear. It might not be that 30% reduction is closer to the goal than a 10% reduction; because the best path from here to the 80% reduction might require three years of very little change while innovation continues, and a reduction might happen very sharply at some point when the new technologies are perfected (see Gates, 2010).

While Brock is ostensibly trying to provide practical ways to solve particular problems for global justice, she gives no attention to how *likely* it is that the proposals she outlines could be brought to pass, or *how hard* it might be to implement them, and only cursory attention to *what the obstacles might be*. While it is relevant that there's a *way* to achieve some outcome, it's surely also relevant what the *chance* of achieving that outcome is. She wants to use feasibility as a weapon against critics of cosmopolitanism, but to do that it must be clear what it is for something to be feasible, and thus clear when and how a critic has actually been defeated.

3.4 Conclusion

In this chapter I have considered some of the ways feasibility is used in the arguments of various prominent political philosophers, in order to demonstrate that the concept plays an important role and that we thus stand to benefit from getting clear about it. I presented some of the elements of a concept of feasibility that have been defended in the literature so far. I argued that some of these do not get us close enough to a workable account (for example, most of Hawthorn's account is too retrospective), while many suggest important elements that should be combined in an explication of the concept, for instance Buchanan's 'feasibility' and 'accessibility', Jensen's 'indirect diachronic abilities', and both Gilabert and Cohen's inclusion of 'stability'. While Buchanan's 'accessibility' and Brock's 'pathways' are important, I have argued that a further element is needed, namely some probability of success. It is not enough that there is some way of getting from *a* to *b*, it matters that the way from *a* to *b* has a good chance of succeeding if we choose to take it. In the next couple of chapters I shall develop a version of feasibility that combines and builds on these important elements. In Chapter 4 I argue for a binary account of feasibility building on discussions of 'ought implies can' in Moral Philosophy, and closest to Buchanan's 'accessibility' and Brock's 'pathways' accounts. In Chapter 5 I argue for a graded account of feasibility that incorporates the probabilistic element I have said in this chapter is important.

Chapter 4

Political Feasibility I: The Binary Sense

4.1 Introduction

The 'ought implies can' project in moral philosophy is structurally similar to the project of seeking feasibility constraints for political theory. Both want facts to play a certain kind of role in limiting theories, or the recommendations of theories. In the first part of this chapter, I will look more closely at the discussion about 'ought implies can' and what is achieved there, to see whether it can get us part of the way toward a satisfying account of political feasibility. I will argue that ultimately the 'ought implies can' discussion gets us only part of the way toward one of the many important roles feasibility can play, namely the binary 'ruling out' role. In the second part of the chapter I will build on that discussion to fully develop the concept of feasibility in that role.

4.2 Ought implies Can

> It has been thought that if "ought" implies "can", and if what a man "can" do can be ascertained by scientific induction, the principle suggests a way of rooting morals and ethics in social science and psychology. Study what men "can" and "cannot" do before you levy "oughts" upon them; otherwise your "oughts" are utopian, uninformed and tragically misleading.[18]
>
> James Ward Smith (1961).

What the 'ought implies can' constraint in moral philosophy tries to capture is the idea that there is something wrong with a theory that prescribes that people do what they cannot do. If a man cannot swim, we should not expect him to rescue a nearby drowning child. If a woman is poor, we should not expect her to make a large donation to charity. And so on. But what exactly is wrong with a theory that requires people to do what they cannot? There are many plausible answers to that question. You might think it's unfair to require people to do what they cannot do, you might think it's irrational (if they can't do it, where's the sense in requiring them to?), or, you might think it's simply pointless (why waste breath telling people what to do when there's no chance that they will go out and do it?) Whatever the reason, it seems natural that we limit people's obligations to those things it is actually possible for them to do. If we perform a simple contraposition on the principle that 'ought implies can', we have the formulation in

18 Smith goes on to argue against the 'ought implies can' principle being used in that way, going on to say 'I fear that there are some self-styled "naturalists" who have used the principle as a bludgeon with which to beat ideals not to their taste'. His own account is that 'cannots' do nothing but reveal further 'oughts'.

terms of constraints: if it is not the case that an agent can φ then it is not the case that an agent ought to φ. (I need hardly point out that this is not the same as the normative claim that an agent ought *not* to φ).

There are two roles which the principle that 'ought implies can' has traditionally been expected to play. The first is in assigning culpability. If someone has the capacity to fulfill an all-things-considered obligation and nonetheless fails to do so, we will be justified in assigning blame or responsibility to them. Inversely, if someone is unable to meet an obligation, then they should not be held to be blameworthy when they fail to meet it. The second role is dissolving obligation, in other words ruling some candidates out of the normative space. Suppose an ethical teleologist proposes a persuasive moral theory in which persons are obliged to care no more about their own children than the children in distant lands and distant futures. If it can be established that people are unable to either cease caring especially for their own children, or begin caring especially for all the children in distant lands and futures, then there is no such obligation.

The second sense is that which is most interesting for this chapter, and the thesis in general. I will bracket the issue of assigning or withholding blame and responsibility in favour of the issue of dissolving obligation (with a brief exception in Chapter 7). The fact that some alleged obligation violates 'ought implies can' should rule that alleged obligation out as actually being obligatory. Likewise in political theory, the fact that some political proposal violates feasibility constraints should function to rule that proposal out from serious consideration for implementation. If we can specify 'ought implies can' for the political, then we'll have a tool with which we can rule out theories claiming that particular things are politically obligatory. (This is to follow the methodology of certain schools of thought in epistemology, for instance Karl Popper's falsificationism, in which refuting some conjecture serves to diminish the set of conjectures taken as known (Popper [1963] 2004); or Jaako Hintikka and Fred Dretske's analyses of epistemic notions like belief, knowledge and information, in terms of different theories of possibility (Hintikka, 1962; Dretske, 1983)).

But are feasibility constraints *just* 'ought implies can' constraints? Let's begin from the assumption that political infeasibility is just a subject-specific subset of one of the primary functions of 'ought implies can', namely its ruling-out function. Then to get a handle on feasibility, we need to first get a handle on 'ought implies can'. Clearly how we interpret the principle that 'ought implies can' rests entirely on how we understand its components, the oughts, the cans, and the kind of implication relation that is supposed

to hold between them. Does the principle claim that *all* oughts imply can? Or only some? And if only some, which ones, and what justifies the demarcation? Does 'can' mean 'is logically possible that…' or 'is physically possible that…' or something more tightly circumscribed? What is the implication from 'ought' to 'can'? Does 'ought' *always* imply can, or just usually? If always, then by the material conditional, or by presupposition, or something else entirely? Questions about the constituent parts of the principle aside, there is the question of whether the principle is descriptive or normative (does it follow from the semantics of 'ought' that 'ought implies can'? Or is it just that we believe 'ought' *ought* to imply 'can'?), and whether the principles violates Hume's law (no 'ought' from an 'is') by deriving an evaluative conclusion from a descriptive premise.

I will touch on most of these issues. The argument I will develop is as follows. Firstly, I will argue that there's no sense in which 'ought' implies 'can' on a completely unrestricted reading of 'ought' (i.e. there are lots of kinds of oughts that don't imply can). Maybe we shouldn't call these 'oughts', or maybe not all oughts imply can; it doesn't matter much. Secondly, I will argue that to preserve its use in ruling-out candidate normative claims the implication from 'ought' to 'can' must be logically primitive. Finally, I will suggest that the utility of the principle hangs on a plausible specification of 'can', in other words, what kinds of inabilities are sufficient to dissolve an 'ought' claim. The latter issue will be most important to the thesis. Just briefly before moving on to develop a version of political feasibility taking its lead from 'ought implies can', I will address the question of whether any formulation of 'ought implies can' violates Hume's law, arguing that it does not.

4.2.1 The 'ought' in 'ought implies can'

In this section, I will survey several of the familiar senses of 'ought' in moral philosophy and ordinary language, with the simple aim of illustrating that there can be no unrestricted reading of ought on which ought implies can. Are there at least some oughts which imply can? Does 'ought implies can' just mean 'oughts of kind *k* imply can?' If so, it will need to be made clear what the relevant kinds of oughts are, and whether non-ideal political philosophy deals in these kinds of oughts (and therefore is properly constrained by the principle).

4.2.1.1 Oughts that require and urge

James Ward Smith distinguishes at least five senses of ought, 'with no pretense at completeness' (Smith, 1961, p. 363). These are oughts of *prediction*, e.g. in response to a question about whether a very dependable friend will show up to a party, "sure, she ought to"; *requirement*, e.g. when a student has an appointment, "he ought to come"; *urging*, e.g. when training an athlete, "you ought to jump even higher tomorrow"; *wishing*, e.g. "he ought to love her back"; and *advising*, e.g. to a student's mother, "your son ought to go to this particular university" (Smith, 1961, p. 363-365). Often, Smith argues, these different senses of ought merge into one another.

Although these are all expressed by way of 'ought', they do seem to fall quite naturally into two categories: *urging* and *requiring*. A defender of 'ought implies can' might argue that 'ought' is meant only in the second sense of requiring. Then while we might utter sentences like 'he ought to love her back', what we're *really* saying is something like 'the world would be better if he loved her back', or 'I really wish he loved her back'. As Richard Joyce points out, the fact that there is a coincidence in one language between words used for moral directives and words used for other things doesn't prove anything, especially when the words are not the same in other languages (Joyce, 2006, Ch. 5). If that is right, then it may well be the case, as Ward Smith in fact argues, that the principle that 'ought implies can' is 'primarily, or even wholly, concerned with the ought of requiring' (Ward Smith, 1961, p. 365).

In fact, other writers have limited the scope of ought in this way before even beginning their discussion of 'ought implies can', e.g. Charles Pigden who focuses exclusively upon imperatives (Pigden, 1990). Gerald Cohen defines principles as 'directives upon agents' before beginning his well-known discussion of whether principles or facts are ultimate, so if oughts are equivalent to principles then he can be taken as limiting the scope of ought in the same way (Cohen, 2003)). Certainly the oughts of requirement look like good candidates for implying can; they are after all concerned with directing action. What is clear is that oughts of urging, whether they are properly regarded as oughts or not, do not seem to be the kind of thing that necessarily have to be able to be done. A coach might coherently urge his athlete to jump higher at the next training even though she is jumping at her physical limit. The world might be considered better if a man were to return the love of his female admirer, even though he cannot (let's say he is already committed, or he is gay, or he is a sociopath). And a

particular university might be best for a student, whether he can get into it or not. So we see already that some 'oughts' look like they should properly be constrained by 'can', while others don't.

4.2.1.2 'Ought to be' and 'ought to do', owned and unowned oughts

John Broome also identifies what look like different senses of 'ought'. He talks about the difference between normative and non-normative oughts (contrast 'you ought not tell lies' with 'the plural of 'mouse' ought to be 'mouses'), insisting that the two senses of ought are sharply distinct. The former are normative, the latter are not. Within the set of normative oughts, he separates 'owned' from 'unowned' oughts, where an ought is owned if it is related to an individual (more formally: is a relation holding between an owner and a proposition), and is unowned if not related to an individual. This distinction between owned and unowned oughts is reminiscent of the more familiar distinction between 'ought to do...' and 'ought to be...'. When we say that something ought to be done, we usually say *who* ought to do it. So the ought is owned. But when we say that something ought to be the case, we might leave the details of how it becomes the case aside, so that the ought is unowned. An example of an owned ought is 'Leone ought to pay off her loan'; an example of an unowned ought is 'it ought to be the case that life is not so unfair' (Broome, forthcoming, Ch. 2).[19]

There is some controversy over whether there are any genuine unowned oughts. Broome himself does not commit to their existence, he merely comments on the semantics of 'ought' in ordinary language. One might argue that all 'oughts' are owned, and it is only that sometimes we do not know who owns them, or that God or some other metaphysical entity owns them, or that they are collectively owned. One might even stipulate that any 'ought' unable to be parsed as owned by an agent isn't a genuine 'ought' at all. So, one strategy is to deny the existence of unowned 'oughts', or explain them in terms of owned 'oughts'. But if there are real unowned oughts ('ought to be's) these obviously do not imply can. That is because if there is no owner whose capacity can be assessed to determine whether the ought is realizable, then such a constraint ceases to make any sense. Think about the claim 'it ought to be that things are a bit more peaceful'. If that doesn't mean that particular individuals ought to be more

19 All page numbers refer to the March 2009 version of Broome's typescript.

peaceful, if it's just something about the environment, then it looks unowned. But then it's not clear what the standard of assessment might be any more. If there's no way of saying whether it can be more peaceful or not, then perhaps unowned 'oughts' too are not of a kind that require empirical constraint.

Ralph Wedgewood (2009) argues that there are two distinct senses of oughts, oughts of practical reasoning, and oughts of desiring. His concern is whether the two senses need a unified semantic treatment, but in the course of the discussion he makes clear that he thinks 'ought' has at least these two distinct senses and is context-sensitive in ordinary language. One way to summarize the difference between the two is to note that while the word 'ought' is crucial to oughts of practical reasoning, e.g. in the proposition 'Wolfgang ought to not brush his teeth for quite so long today', it can be substituted without loss of meaning in the case of oughts of desiring, e.g. in the proposition 'it ought to be the case that the sun shines tomorrow', which can be traded for 'it will be good if the sun shines tomorrow' or something similar. Wedgewood argues that 'ought' is an operator on propositions, and that oughts of practical reasoning are implicitly indexed to an agent and a time, while oughts of desiring are not so indexed (Wedgewood, 2009, esp. Ch. 4, Sec. 4.3).

4.2.1.3 Conflicting oughts?

Within the subset of owned oughts, Broome identifies apparently conflicting oughts: rational, prudential and moral among them. I might be morally required to do the action in any given situation which has the highest chance of bringing about the best consequences, prudentially required to instead formulate general rules and adhere to them, because I have neither the time nor the mental capacity to reason through each particular situation, and rationally required to refuse to be guided by a system of morality which cares only for consequences. But that's a problem for 'ought implies can': if there are conflicting oughts, then I can't possibly fulfill all of their requirements; in which case, there are 'oughts' where there are 'cannots'.

What look like conflicting oughts do not lead Broome to deny that there is a central concept of ought. He says that moral, rational and prudential oughts are *adverbial*, and would do better if rephrased as requirements. The central concept of ought is ‘ought, all things considered’, which means that we must weigh the competing requirements against one another to reach the ultimate ought:

> Here is the picture I have painted of the structure of normativity. There are various sources of requirements. The requirements that issue from some of these sources are normative. Separate normative requirements, issuing from different sources, feed into a central, overall ought; together they determine what you ought to do, ought to believe, ought to be, and so on … Different sources of requirements do not threaten the single central normative concept of *ought* (Broome, forthcoming, p. 26).

If one is happy to accept such a thing as an all things considered ought, then this solution to the proliferating oughts problem looks like a good way to go. If there is no conflict, then there are no 'oughts' where there are 'cannots'. The apparently different senses of ought identified by Broome (and some of those identified by Ward Smith, discussed in Section 4.2.1.2) are in fact only sources of requirement that inform an overall ought of practical reasoning. And surely the ought of practical reasoning will have to be constrained by 'can', given as it is directly concerned with what we should do.

But there is a great deal of scepticism as to whether the requirements of morality and requirements of prudence, in particular, can ever be reconciled. Many think they just give conflicting answers about what to do, so that there will be no interesting sense in which the all-things-considered ought combines their judgements and tells us what to do (see e.g. Sidgewick, 1874*;* Crisp, 2006; Parfit, forthcoming). If that worry is well-founded, then we are left with competing 'oughts' (or 'sources of requirement') which cannot be reconciled, and in that case it really does seem that there are 'oughts' where there are cannots, because the 'cannots' are simply a feature of there being so many competing 'oughts', not all of which can be done. This might be the best way to view situations of tragic conflict. Rather than saying there's only one thing that ought to be done, we might say there are two (or more) things that genuinely ought to be done; it's just a sad fact of the situation that not all of them can be done. This does seem more intuitive than insisting that 'ought' always implies can, which would mean having to agree that in the tragic conflict situation, one of the choices that looks morally obligatory actually for some complicated reasons to do with one's moral theory is not.

4.2.1.4 Oughts that attribute blame

Peter Vranas, in a modern attempt at formulating 'ought implies can', describes a case in which one agent says to another: 'you ought to feel grateful to her' (Vranas, 2007). In the case, the agent being spoken to does not feel grateful, and perhaps cannot, due to not being the kind of person who expresses gratefulness when it is appropriate. Cases designed to show that an obligation stands even in the face of its unrealizability

are stock-standard in the 'ought implies can' literature, but Vranas uses it not to attack the principle but rather to make the following point: rather than ascribing obligation, the second agent might be ascribing blameworthiness. Perhaps the meaning of the 'ought' in that case is to say that the agent ought to have, *in the past*, cultivated appropriate responses such that she would feel gratitude where it was deserved. To say that 'oughts' of blaming are not constrained by 'can' might seem counterintuitive at first: surely we think someone's ability to do something is highly relevant to whether she can be found blameworthy for not doing it.

That is exactly right, so long as we separate synchronic from diachronic ability. Notice the difference between '*x* ought to have never happened', which is a synchronic ought about a past which cannot be changed, and 'ought never to have happened', which is actually ambiguous but can be used to denote a diachronic ought about a counterfactual, which says that things ought, in the past, to have gone differently. 'Oughts' of blaming can simultaneously acknowledge that an action is synchronically impossible, but was diachronically possible. 'Oughts' of blaming are 'oughts' with temporally removed 'cans'. Whether we take this as further proof that 'ought implies can' cannot use an unrestricted 'ought' will depend on whether we think the 'can' is in general synchronic or diachronic, but in standard cases it seems to be synchronic, and thus on a standard understanding 'oughts' of blaming are a further example of 'oughts' not constrained by can (e.g. can-now), even though they do seem to be constrained by can in the diachronic sense (e.g. could have, then).

4.2.1.5 Oughts that express a judgement

Ascribing blame is not the only function of a synchronically impossible 'ought'. Another function is to express a judgement, to comment, perhaps, that a state of affairs would be good or fitting in the given circumstances, even though it is by now impossible to bring it about (Vranas, 2007, p. 6). This is similar to the interpretation of unowned oughts or 'ought to be'. When people utter propositions like 'it ought to be the case that life is not so unfair', they are saying that *it would be better if* life was fairer. The 'oughts' are not 'oughts' in the sense of requirements upon individuals to *do* certain things, but rather comparative judgements, in which the ought could as well be traded in for 'it would be better if' or 'it would be good if'. But why should my judgement about better worlds be constrained by the way the world happens to be? In fact, it seems that it clearly should

not be so constrained, for then theorizing about value itself would be imaginatively constrained, and yet there is a clear role for the imagination in such theory. Pablo Gilabert, for example, points out that creative thinking about what is desirable might inspire people to investigate whether those things are possible (Gilabert, ms., 2008). But clearly people cannot know in advance of finding out whether or not something is possible that it is possible, and thus an appropriate expression of judgement.

4.2.1.6 Oughts in ideal theory

The final challenge to any unrestricted reading of 'ought implies can', already discussed in Chapter 2, is that if it were true that ought implies can unrestrictedly, then there would be no role for ideal theory. Ideal theory is a part of normative theory which contains idealizing assumptions about the real world, e.g. does not take 'can' seriously in the same way that non-ideal theory does. I argued in Chapter 2 that ideal theory fulfills several valuable functions, first by giving us valuable insight into our normative concepts, and second by providing a normative standard which can be more or less relaxed. Others have argued that ideal theory and non-ideal theory are part of a common project (Stemplowska, 2008). Those considerations give us good reason to deny that 'ought' implies 'can' (and also the normative reading of the principle, 'ought *ought* to imply can') unrestrictedly, or in other words, that *all* oughts imply can.

4.2.1.7 Multiplicities of oughts

In this section I have surveyed several different kinds of oughts, oughts that require and urge, owned and unowned oughts, 'ought to be' and 'ought to do', oughts that attribute blame, conflicting oughts including the all things considered ought of practical reason, oughts expressing judgement, and oughts in ideal theory. The fact that there look to be many different kinds of oughts, not all of which imply 'can', does not immediately entail that 'ought implies can' is false. First of all, we might insist that all oughts imply can, but say that the 'can' is weaker in some cases and stronger in others. Perhaps oughts in ideal theory are constrained by only the most general facts about human existence, while 'ought to do' is constrained by more contingent and context-specific facts. In fact, I suggested something along exactly these lines in Section 2.4.

Second, we might just deny that all of the senses of ought just mentioned are

genuine oughts. Perhaps oughts that are constrained by can are the only genuine oughts, and everything else is evaluative rather than normative, or about hoping and wishing rather than prescribing. That is not implausible; some people use 'normative' to mean 'action-guiding', and thus would fail to grant to evaluative oughts the kind of normativity that is the standard concern of moral and political philosophers.

Third and finally, we might prefer to say that indeed, there are many kinds of oughts, and not all of them imply can. That tells us that 'ought implies can' is false *if* it is meant to include all kinds of oughts within its scope. But perhaps it is not meant to do that; perhaps what is more important is that we isolate the particular kinds of oughts that should properly be constrained by can. Gricean maxims of conversation can do a lot of the work in justifying a principle along those lines. They include things like 'be concise', 'be relevant', and 'be informative', which together yield something close to 'do not waste your breath by saying pointless things' (Grice, 1989). Such maxims suggest that only imperatives, prescriptions, directives, and in general normative claims intended to be action-guiding, should be assumed to be constrained by can. Then the best understanding of 'ought implies can' would be something like 'oughts intended to be action-guiding imply can' (which was roughly the choice of Chapter 2, to limit feasibility constraints to non-ideal theory).

What is nice about this third option is that surely in non-ideal political theory we have one of the strongest cases of oughts intended to be action-guiding, and therefore oughts that should be constrained by cans. We would spend taxpayer money on reforms, ask people to put effort into the changes we want to make, we would ask people to trust one another (because people are unlikely to act if they think they'll be the only one), in many cases we would invest public resources in projects, and stake our careers and reputations on their succeeding. For all of these reasons, political oughts (at least those involved in public policy-making), and oughts in non-ideal political theory, look like particularly strong candidates for restriction by 'ought implies can'. So whether the principle applies to all oughts but with varying strengths of 'can', or only to some oughts with a uniform strength of 'can', it surely applies to political oughts if it applies to any.[20] I shall simply leave it as an open question which of the three options is to be preferred in dealing with the multiplicity of apparent oughts.

20 This is not to say that political philosophy always deals in oughts that are meant to be action-guiding. Sometimes political theories are utopian or ideal, and in that case we might want to say that their prescriptions are only evaluative, or only conditionally normative (e.g. if certain pre-conditions can be obtained). If and when they are intended to be action-guiding, they will fall into the same category as the genuinely political oughts mentioned here.

4.2.2 Entailment from 'ought' to 'can'

Most philosophers have taken the claim that 'ought implies can' to be descriptive. Some take it to be a logical truth. There are other possibilities, though. One is that the implication from ought to can works by way of pragmatic conversational implicature. There are norms against saying pointless things, so if you say that I ought to do something, you should generally believe that I can (Sinnott-Armstrong, 1984). Another is that it works by way of presupposition. The statement 'you ought to *x*' presupposes the truth of 'you can *x*', just as the statement 'the King of France is bald' presupposes the truth of 'there is a King of France' (on semantic and pragmatic presupposition in general see Stalnaker, 1973; Stalnaker, 2002; and on its application to 'ought implies can' see Collingridge, 1977). Yet another is that it works via some sort of shared moral belief. We believe as common ground that in a fair society, people are not obliged to do what they cannot do (Pigden, 1990; Collingridge, 1977). (For a recent survey of these possibilities see (Vranas, 2007)).

Despite these many possibilities, only one, logical entailment,[21] is a serious contender, given the function we want 'ought implies can' to perform, i.e. contraposition. Philosophers have been much more interested in the contraposed claim that 'not-can implies not-ought' than in the principle in its standard form:

> Proponents [of 'ought implies can'] generally see the relation in question as one which allows contraposition, since they see the function of the thesis as ruling out ought-judgements about what cannot be done (Collingridge, 1977, p. 349).

If the implication from 'ought' to 'can' works by presupposition, or conversational implicature, or moral implication, then the contraposition 'not-can implies not-ought' will not be valid. That's because those other kinds of implication allow that 'ought' and 'cannot' are compossible. The problem is that there are abundant counterexamples to the idea that the entailment is logical. To give just one example, Charles Pigden presents a case in which a doctor tells his patient to do something the patient clearly cannot do. It may not be *nice* for the doctor to do so, he admits, but it might serve a useful purpose, namely, to 'publish the powerlessness of the addressee and bring his incapacity forcibly to his notice' (Pigden, 1990, p. 11). Telling a smoker that he ought to stop smoking when he cannot may force him to realize his addiction. Pigden's claim is that there are cases

21 The logical entailment probably has to be semantic. The material and indicative conditionals also allow contraposition, but semantic entailment is by far the most plausible and best-defended. As we shall soon see though, we don't need to worry about the details of the entailment.

where ‘it is intelligible, appropriate, and not logically odd to command what can’t be done’ (Pigden, 1990, p. 10). And those kinds of counterexamples, of which there are many, are sufficient to undermine the idea that the relation between ought and can could be logical entailment.

However, having already noted that not all oughts imply can, it should not be surprising to find that there are counterexamples to the logical entailment between 'ought' and 'can' that utilize some of the 'oughts' I have already said don't necessarily imply 'can'. A genuinely damaging counterexample would have to show that even for the most central cases of 'ought', e.g. those which are intended by the utterer to act as directives or prescriptions, there is no necessary implication of 'can'. Defenders of 'ought implies can' will simply deal with examples like Pigden's by saying that the utterer did not genuinely intend his ought to be action-guiding; or at least not in the way the surface grammar suggests. Rather, the doctor in Pigden's example intended his prescription to cause the addressee to realize an addiction. The statement was action-guiding and satisfied 'ought implies can' in that the intended action was realizable (it would be irrational for the doctor to use it as a means to the patient's realization of his own addiction if he thought the patient was incapable of realizing his addiction), but it was not intended to be action-guiding, and thus doesn't violate 'ought implies can', with respect to its surface grammar (i.e. that the patient really should stop smoking). So contraposition by way of logical entailment stands so long as we are careful about the kinds of oughts that fall under the scope of the principle.

In any case, the question of what kind of entailment takes us from ought to can might fairly be seen as beside the point. I mentioned above that philosophers have been primarily concerned with 'ought implies can' in its contraposed form. But then why not just start with a principle of that form? One possibility is 'inability entails non-requirement'.[22] This is a plausible principle, and it captures what most people have been interested in when discussing 'ought implies can' all along. Certainly it is not the original formulation proposed by Kant. But if it is something on which we can in general agree, then we might just take it as a starting assumption, rather than working to achieve a contraposition on the older and more-discussed principle that 'ought implies can'. The important questions still remain, namely, 'what counts for inability?' ('what does 'can' mean?'), and 'what kinds of requirement are subject to this constraint?' ('what kinds of

22 I am grateful to David Estlund for this point, and for his comments on my paper “Why Ought Implies Can is Not a Good Way to Approach the Feasibility Issue”, presented at the Australian National University Workshop in Political Feasibility in 2008.

'oughts' are properly constrained by 'can'?') We have already dealt with the question of requirement (ought), now it is time to turn our attention to the question of ability (can).

4.2.3 'Can'

David Lewis has argued with respect to time travel that there's a perfectly meaningful sense in which a person 'can' travel back to the past and kill their own grandfather, and a perfectly meaningful sense in which they 'cannot'. The explanation for this seeming contradiction is that the word 'can' is equivocal between different senses (Lewis, 1976, p. 6). The time-traveller can kill his own grandfather, in the sense that he 'has what it takes' (following Vranas we can say he has the ability plus the opportunity), but he cannot kill his grandfather, in the sense that it is logically impossible to change the past. Context usually tells us which delineation of 'can' we are using, but it may not determine it perfectly. The lesson is that 'can' (and correspondingly 'cannot') is not fixed in its meaning, so we must always be explicit about the delineation of relevant facts we have in mind to establish it.

Bart Streumer offers a *tensed* account of 'ought implies can'.[23] He differentiates three senses of 'can', roughly equivalent to past, present and future: 'was able to', 'is able to', and 'will be able to' (Streumer, 2003, p. 219-228). This way of dealing with 'can' avoids the concern some have had (e.g. Sinnott-Armstrong, 1984) that agents might escape moral obligation by *making* themselves unable to fulfill them. On Streumer's tensed account:

> ...though it is no longer true after 4.55 that Adams can meet Brown at 6.00, it is still true after 4.55 that Adams *could have met* Brown at 6.00. … So if we formulate the view that 'ought' entails 'can' in a tensed way, we can still blame Adams for not *having met* Brown at 6.00. And that seems exactly right (Streumer, 2003, p. 225).

For political purposes, we will most often be concerned with future-oriented ability, and dissolving future-oriented prescriptions for reasons that they cannot be realized. We are not much concerned with attributing blame and responsibility, so we can leave at least the past sense of ability behind. What is important is the present and future sense: what a person can do now, and what she will be able to do in the future.

Steve Sapontzis separates situation-specific impossibilities from deep impossibilities (Sapontzis, 1991). Bad weather might bring it about that I cannot fly from

23 In fact I think this just formalizes a point that most people make quite naturally.

Sydney to Melbourne, and will therefore miss an important conference. But deep physiological constraints are what make it the case that I cannot myself fly (e.g. I am not a biological organism capable of flying). On the same note we might distinguish timeless from time-indexed impossibilities; a prescription indexed to a specific time must take seriously the known constraints of that time, while a timeless prescription must take seriously the known constraints across all times.

Vranas argues against equating 'can' with possibility and in favour of equating it instead with potentiality, using the example of a student having an obligation to hand in a paper by nine a.m. In this case the distinction between possible actions and potential actions becomes sharper: at one minute to nine it is *possible* that the student begins typing at superhuman speed and turns the essay in on time, but turning in the essay doesn't look to be a *potential* action of the student's. Vranas dismisses the former possibility as 'exotic', arguing that it should have little bearing on our considerations about obligation. If the 'can' of 'ought implies can' (or the 'inability' of 'inability entails non-requirement') is only something like conceptual possibility, then it fails to achieve its function as a ruler-out of unachievable oughts. The obvious problem, then, is to find an appropriate specification of 'can', one that is not inconsistent with the intention of those who wield the principle. Vranas himself goes for a formulation in which 'can' is equivalent to whatever is a potential action of an agent, where the necessary conditions for potential actions are *ability* and *opportunity* (Vranas, 2007).

The problem with this account is that an agent has many potential actions, especially across time. What should we say about the father who consistently puts off having his car tires replaced, in spite of their increasing baldness, and ends up killing several people in a terrible car accident? There were many points at which changing his tires was a potential action of his, even if at the time of the accident losing control of the car was outside of his control. Or what of the bystander who cannot provide aid in a medical emergency, because he failed in the past to obtain the appropriate training? Obtaining medical training was a potential action of his at many points in his life, and it is the decision not to have obtained it that is causally responsible for his inability to provide aid when it is needed. Any account of inability sufficient to dissolve requirement must also specify an account of the abilities it is reasonable for a person *not* to have developed. If not, the account will be both overly demanding (requiring that people spend all of their time getting themselves into a position to realize potential future obligations), and implausibly radical in its attribution of blame and responsibility,

maintaining that people are guilty for not having done all that they could have to be able to fulfill a future *prima facie* requirement.

As a final word on 'inability' (cannot), in the last chapter I introduced Jensen's account of practical possibility, in which he suggested that of the four necessary and sufficient conditions, logical consistency, non-violation of laws of nature, fixed history of the world, and natural human ability, the latter is the most crucial. He separates natural human ability in the paper into three categories: synchronic, direct diachronic, and indirect diachronic. Synchronic and direct diachronic abilities are transparent to their authors and others, so the most important kind of ability for us to figure out, as philosophers, is indirect diachronic ability. The distinction applies to individuals as well as to groups (Jensen, 2009).

What's nice about Jensen's account is that it makes some useful distinctions for how we think about 'ability'. It points out that for something to be practically possible it doesn't have to be the case that it is possible *now*, it just has to be the case that it's possible *at some point*. The introduction of indirect diachronic abilities makes precise the idea that we can do something now to make it the case that we can do something else later on; and so on *ad infinitum*. What remains to be seen is how we are to establish that those kinds of abilities are present (on which more in the coming chapters).

4.2.4 An objection: is 'ought implies can' valid, given Hume's law?

Whether we talk in terms of 'ought implies can' or in terms of the more direct principle that 'inability entails non-requirement', we have to face the challenge posed by David Hume that no 'ought' can be validly derived from an 'is' (hereafter 'Is/Ought'). That is to say, no normative or evaluative conclusions can be derived from descriptive premises. This fairly uncontroversial claim is often taken to contradict 'ought implies can'. The tension comes from the fact that Is/Ought dictates that no moral conclusions come from purely non-moral premises, while 'ought implies can' uses non-moral facts, in this case empirical facts about what an agent cannot do, to produce conclusions which are sometimes moral.

I say 'sometimes moral' rather than simply 'moral' because as mentioned already, clearly 'it is not the case that one ought to *A*' is not the same as 'it is the case that one ought not to *A*'. On the standard framework familiar from deontic logic, there are only three modal operators. Something can be obligatory, prohibited, or permissible. In

ordinary language, 'it ought to be that *A*' corresponds to '*A* is obligatory', 'it ought not to be that *A*' corresponds to '*A* is prohibited', and 'it's not the case that it ought to be that *A*' corresponds to '*A* is permissible'. The latter is what we get from a contraposition of 'ought implies can'. And that statement is ambiguous between two of the moral categories, permissibility and prohibition. That means that in some cases, 'ought implies can' doesn't by contraposition yield a straightforwardly normative conclusion, and so doesn't violate Hume's Law. It merely says that it's not that case that *A* ought to be done (*A* is not obligatory), leaving it open whether *A* ought not to be done, or it is permissible that *A* be done.

But imagine that there are only two options in some scenario: save the drowning baby, or do not save the drowning baby (because any action other than those sufficient for saving it will count as not saving it, e.g. my walking on by, listening to music, dancing a little, and so on). If 'ought implies can' shows that I cannot save the drowning baby, perhaps because I cannot swim, then by contraposition it is not the case that I ought to save the drowning baby, i.e. not the case that *A* is obligatory. But the two statements 'it is not the case that you ought to save the drowning baby' and 'you ought to save the drowning baby' are contradictory, which rules out the latter, leaving the only other option on the table 'you ought not to save the drowning baby'. This is not a straightforward derivation, rather it is that the contraposition of 'ought implies can' sometimes justifies choosing one action over another when options are limited. And in those cases, 'ought implies can' does take normative conclusions from descriptive premises (although notice that the normativity had to be there in the first place), and so does violate Hume's law. Let's see whether we can resolve the challenge, for that special subset of cases.

There are a few unsatisfying attempts in the literature to dissolve Is/Ought in the face of 'ought implies can'. A standard formulation of Is/Ought is 'no valid argument has a conclusion that is a moral claim and premises that form a consistent set of non-moral claims' (Vranas, 2007).[24] One strategy of arguing against Is/Ought, or at least against this formulation of it, has been to utilize one of the basic rules of first order predicate logic, namely disjunction introduction. This allows the valid move from any non-moral premise to the disjunction of that premise with some moral claim, any example of which is a counterexample to Is/Ought. For instance:

24 This formulation is from (Vranas, 2007), but he refers to similar formulations in (Brink, 1989, p. 146) (Harrison, 1967, p. 70), (Prior, 1960, p. 199-201), (Rynin, 1957, p. 308), (Schurz, 1991, p. 38) and also (Searle, 1964, p. 43).

1) My new gumboots are yellow

My new gumboots are yellow, *or* killing kittens for fun is wrong.

There we have a valid argument from an 'is' premise to an 'ought' conclusion, which provides a counterexample to Is/Ought. One could argue that the disjunct in the conclusion is actually non-moral, rather than moral, but then the *modus tollens* argument from the disjunction and the negation of the premise still suffices as a counterexample to Is/Ought) (Vranas, 2007, p. 15). This is not a crushing blow for the defender of Is/Ought, however. It just serves to show that he might need to exclude arguments with disjunctive conclusions from the scope of his claim.

A different strategy of rejecting Is/Ought has been to use another of the basic rules of logic, namely that which allows the existential innocence of universal quantification (if no *A*s exist, then it is true that all *A*s are *B*s). For example:

1) No kittens exist

No kitten ought to be killed for fun

Again, we have a valid argument from a non-moral premise to a moral conclusion (Prior, 1960, p. 202). As with the example above, this example suffices to defeat Is/Ought, but probably the real lesson is that the defender of Is/Ought needs to be precise about the kind of derivation he has in mind. One thing to notice with these counterexamples that utilize permissible moves in first order predicate logic is that the arguments target *sentences* containing moral and non-moral propositions, rather than *worlds* containing moral and / or non-moral facts. Are the same kinds of escape routes open to the person who would defend 'ought implies can' against the Is/Ought challenge when we talk in terms of worlds?[25]

What is at issue is the relation between moral facts and non-moral facts, between alleged obligations and empirical constraints. So instead of asking about the validity and soundness of arguments moving from non-moral premises to moral conclusions, i.e. framing our discussion in terms of sentences, we can rather ask about worlds. Could there be a possible world exactly identical to ours with regard to the 'is'-facts, the facts about how things are, that differed in 'ought'-facts, the facts about how things ought to

25 I am grateful to Wolfgang Schwarz for the suggestion to talk in terms of worlds instead of sentences, and for helpful discussion with respect to this section of the chapter.

be, or what is morally right and good? If our answer to this question is that there could be no such difference, then we hold one of the following meta-ethical positions: naturalism, nihilism, or what I will call 'necessity'. Naturalists hold that moral facts come out of non-moral facts in some special way, for example Jonathan Dancy (1993) who maintains in his book *Practical Reality* that the good is just whatever we have reason to do, or Frank Jackson's (1998) supervenience thesis. Nihilists hold that there are no moral facts e.g. Charles Pigden (1991; 2007; 2009) with his so-called 'reluctant nihilism' which is a kind of error theory, or Richard Joyce (2006) with his borderline error theoretic claim that evolution neither supports nor completely debunks the existence of independent moral facts. Necessitarians, if there are any, hold that all moral truths are logically or conceptually necessary. I shall refer to the disjunction of these three meta-ethical positions, naturalism, nihilism and necessity, in the rest of this section as 'NNN'. For any of these positions, no worlds identical in non-moral facts will differ with respect to moral facts, and thus there is no conceptual separation between 'is' and 'ought' as the proponents of Is/Ought have maintained.

What is the alternative to the NNN disjunction? Instead of maintaining that there could be no moral difference between worlds identical with respect to non-moral facts, we'd have to maintain that there *could* be such a difference. That would mean accepting that there are possible worlds identical to ours with respect to all the non-moral facts, but differing with respect to the moral facts. This kind of answer runs counter to NNN, so hereafter I refer to it as anti-NNN. The position can be specified in two ways, one stronger than the other. Under weak anti-NNN, which is supported by the open-question argument due to G.E. Moore (1903), 'is'-facts can be taken to rule out several possibilities with regard to 'ought'-facts, while still leaving it open exactly which set of ought facts correspond to the set of 'is'-facts. To restate, the weak version allows that once all the facts about what an agent cannot do are specified, there is still more than one possible way the 'ought'-facts could be. Under strong anti-NNN, on the other hand, it is *completely open* which 'ought'-facts correspond to the given world. Only the strong version contradicts 'ought implies can', because even given all the facts about what agents cannot do, it remains completely open whether they ought morally to do them anyway. That is to say, value remains undefined. A useful way to conceptualize the difference between these positions is numerically. Given the complete set of 'is'-facts, how many ways could the 'ought'-facts be? NNN says there is exactly one way (or more accurately, naturalism and necessity say one, nihilism says none because there are no

'ought'-facts), weak anti-NNN says there is some limited number of ways (some possibilities have been ruled out), and strong anti-NNN says that there are *infinitely* many ways.

Under NNN, 'ought implies can' is not at risk from Is/Ought, because none of naturalism, nihilism or moral necessity have trouble with 'ought'-facts following from 'is'-facts – there is exactly one way the world could be with regard to 'ought'-facts when all the 'is'-facts have been given. Under weak anti-NNN, 'ought implies can' is *still* not at risk from Is/Ought, because that position is consistent with there being *some* 'oughts' which come out of 'is'. The only position in which Is/Ought threatens 'ought implies can' is under strong anti-NNN, the extreme position that leaves it completely open, even given a complete specification of the 'is'-facts, what the 'ought'-facts of some possible world will be. But strong anti-NNN positions are not particularly prevalent in the meta-ethical literature. One example of such a position might be a non-cognitivist version of expressivism which ends up entailing a radical form of moral relativism, but many writers now take expressivism to be a kind of realism, e.g. Simon Blackurn (1993). In what follows I will simply assume that for the meta-ethical reasons just given, the Is/Ought challenge is not strong enough to completely derail inquiry into 'ought implies can' (or the contraposed version of the principle).

4.2.5 Summary: from 'ought implies can' to feasibility

The 'ought implies can' project is first and foremost binary. It uses 'can' as a tool with which to dissolve alleged obligations. A violation of 'can' constraints *rules out* an obligation from having any normative force upon an agent. The first thing to notice is that the principle standardly targets the *actions* of *individuals*. That means it has a different subject matter from feasibility constraints, which I mean to target *outcomes*, recommended by theories, to be brought about by both individuals and *collectives*. The second thing to notice is that 'ought implies can' plays only one role that is structurally similar to those that have been proposed for feasibility, namely the ruling out role. 'Ought implies can' provides a fairly blunt tool. It lets us rule out the proposition that an individual is obliged to act, when it can be shown that the individual cannot act. And in one of its roles, feasibility constraints should rule out the proposition that it is obligatory to pursue or bring about some outcome, because that outcome cannot be realized. But 'ought implies can' does not get us an account of all the interesting roles feasibility plays.

For example, no one has proposed that 'the more a person can, the more they ought' (although this does seem to have some normative bite).

For the rest of this chapter, I will focus on building upon the discussion of 'ought implies can' to develop a structurally similar, binary sense of feasibility. This sense will allow the ruling out of theories or their recommendations when they cannot be implemented. In Chapter 5 I shall turn to the graded sense of feasibility, which will allow comparisons between various theories and sets of recommendations.

4.3 From 'ought implies can' to the binary sense of feasibility

In Berlin, there is a group of people who call themselves 'Carrot Mob'. They want local businesses to switch to using more energy efficient appliances. They do this by approaching business owners and proposing that if they can get many more customers than normal into the store, the owners use profits above the normal amount to fund switching from their current appliances to more energy efficient ones (refrigerators, microwaves, washing machines, freezers and so on).[26] The group takes its name in reference to the 'sticks and carrots' approach to motivation. It chooses to use an incentive rather than a disincentive: the profits are money they would not otherwise have earned, and in addition, more people going to the business means more people knowing of the business, which gives the cafe a chance for repeat custom. Imagine that Carrot Mob's current target is a small cafe in Kreuzkölln called *Krawall & Remidemi*. Now consider the proposition: 'It's feasible that *Krawall & Remidemi* switch to more energy efficient appliances'. This is a claim about an *outcome*, namely that it is one that can be brought about.

In ordinary language we make statements like that all the time. 'It is feasible that *P*', we say, or, 'it is infeasible that *P*'. But what does that actually mean? *When* is an outcome feasible? In the formulations to follow, I will talk about the recommendations of theories, but that should be understood to extend to whole theories as well. Let's start with the following:

(a) The recommendations of some theory are feasible *iff* they can be brought about.

26 See e.g. http://berlin.carrotmob.de/ (and) http://berlin.carrotmob.org/ accessed 01/09/10.

This is a place to start, but much more needs to be said. What does it take for it to be true that the recommendations of some theory can be brought about? It might help to know *whose* actions are being assessed. Thus something like:

(b) The recommendations of some theory are feasible *iff* there exists an agent who can bring them about.

This cannot be exactly right. It's helpful to fix the feasibility of outcomes to particular agents, because then we know who to look at to establish whether the outcome can be brought about. Say we want to know whether it's feasible that the apples be picked from this tree before they fall off and go rotten. Then (b) seems to get things right. That outcome is feasible if some agent exists who can pick the apples from the tree before they fall off. But now suppose we want to know whether it's feasible that all the apples in the *orchard* be picked from their trees before they fall off and go rotten. And suppose that no single individual can bring this outcome about. That isn't a reason to say the outcome is infeasible, because there might be many individuals who could pick the apples, or there might be a collective agent (say, the apple pickers' union) who could bring that outcome about. So I should stipulate that 'agent' can mean an individual or a collective, and that there must be either an agent on this understanding or a *set of agents* who could bring the outcome about. Thus:

(c) The recommendations of some theory are feasible *iff* there exists an agent, or set of agents, who can bring them about.

This is better, but we still need to be more precise about what it means to say that an agent 'can bring about' the outcome that the recommendations of some theory are realized. One way to fill out that detail is to say that the agent must have as one of her *options* an action that would bring about that outcome. Then we can direct our attention to how to say when an agent has an action in her option set, and when she does not. One outcome of an action is always that the action is done. Outcomes include actions, but they are not limited to them (Broome, 1991). Thus:

(d) The recommendations of some theory are feasible *iff* there exists an agent, or set of agents, who has available to her some action or set of actions that could

bring the recommendations about.

Does this get things roughly right, supposing that we can tell a convincing story about which actions are in an agent's option set, and which are not? Returning to the simple apple-picking case, imagine that there is some theory which issues the recommendation that all the apples in the orchard be picked. Perhaps the theory is a broadly consequentialist one, and the obligation is based on the idea that there would be less suffering in a world where the apples were used as food than in the world where the apples were left to rot. To figure out whether the theory's recommendations are actually obligatory, which is to say, whether feasibility in its ruling out role can reject the alleged obligation, we ask whether there exists some agent, or some set of agents, who has the action 'picking all the apples' in her option set. I have allowed for both individual and collective agents, single agents and sets of agents. We simply look to the context and see whether such agents exist. If there is only one individual who could get to the apples in time, and she couldn't pick *all* the apples, then the recommendation is infeasible. If there are many agents who could get to the apples in time, and they could together pick all the apples, then the recommendation is feasible. Likewise if there is a collective agent, say the apple-pickers' union, and it could pick the apples in time, then the recommendation is feasible.

One question we might want to ask at this point is what the spatial and temporal index must be. I have said that we can establish feasibility by looking at which agents exist, and what their options are. But so far that is to ignore spatial and temporal constraints. Is the recommendation that all apples in the orchard be picked really feasible just because a group of expatriate New Zealanders living in Singapore has as one of its options flying home to the South Island orchard to pick all the apples? Or what if there's no agent *now* who could pick the apples, but there will be one in the future, or, interestingly, could be one in the future, because current agents could come together to constitute a collective agent with that action in its option set?

The answer to that question according to (d) is 'yes', as much as it might appear to warrant resistance at first glance. The fact that the expats are in Singapore might make it less likely that they *will* go home to pick the apples, but if they have the option to do so, then it's not ruled out that the apples get picked. Likewise the fact that current agents must come together to constitute a collective agent with some action as an option might make that action less likely to be done, but the fact that current agents could bring

into existence an agent who has it as an option means that the apples being picked is not ruled out.

Feasibility assessments must be indexed to a time. Sometimes the temporal period will be implicit in the alleged obligation, as it is in the apple-picking case ('before they fall off and go rotten'). Sometimes it won't, so we will have to be explicit about it. For example, take the proposition 'it is obligatory that the group of rich countries eliminates poverty in developing countries'. Is that proposition true? It is true if a good moral theory recommends it and if the outcome it cites is feasible. But the proposition lacks a temporal index. It would be odd to say that this recommendation is infeasible because there's no agent that exists *now* who could eliminate poverty in developing countries. It would be odd because there may well be agents who have as one of their options the prior action of coming together to constitute a collective agent (the 'group of rich countries'), and if the agents each chose that action then the requisite collective agent would be brought into existence. Then we might like to say that the proposition is true, and the recommendation contained therein is feasible, because if we take an extended time period as the relevant index, rather than taking it to be 'now', there *is* an agent who has as one of its options fulfilling the recommendation of the theory. The temporal scope matters a lot in giving a definite answer about what is feasible. So a further modification is needed:

> (e) The recommendations of some theory are feasible *iff* there exists an agent, or set of agents, within the designated period of time, who has available to her some action or set of actions[27] that could bring the recommendations about.

What this implies is that when a temporal index is not specified, it will be much harder to rule the recommendations of a theory out as infeasible. Only if no agent could ever exist who would have as an option fulfilling the recommendations of a theory would feasibility be able to rule those recommendations out. It will be much easier to rule out recommendations which are temporally indexed, because then we only have to look at which agents exist, or could be brought into existence, within that time period.

I say we might judge a theory feasible when an agent could be brought to exist

27 For a logic of feasibility to be possible, we must include options of the form 'be such that...'. That is to allow statements like 'it is feasible for Adisyn to be such that 2 + 2 = 4', or 'it is feasible for Olivia to be such that it is raining'. Because these are true, they are actual, and because they are actual, they are *a fortiori* feasible. So in formulating feasibility one has to allow for them, even though they will not be the kinds of things we usually ask feasibility questions about.

who could realize the theory's recommendations within the given temporal period. That means solving global poverty will be feasible because current agents can could come together to constitute 'the global society', which would have as one of its options the global redistribution of its wealth.

But we may often want to focus our feasibility assessments more narrowly than that. In that case we restrict the scope of an assessment before we begin. For example, I might want to know whether cooperation between states to cut carbon emissions to one third of 2010 levels by 2050 is feasible. And I might reason that we just can't count on China to agree to anything, or if they do agree, to follow through on that agreement. If I reason in that way, I will not consider the options of collective agents that include China as members to be the determinants of feasibility. I will restrict the scope of my assessment to the options of agents who I'm sure won't stonewall. So for example, I will look at which other agents there are, e.g. which states, and what they have as options. If the outcome can't be realized without China I will declare it infeasible. If it can be realized without China, I will declare it feasible. Feasibility allows us to make useful assessments, but we get out what we put in. If we are maximally permissive about what counts as an agent we will get maximally permissive answers about what is feasible, and if we are maximally restrictive about what counts as an agent, we will get maximally restrictive answers about what is feasible (and the same goes for how permissive we are about the temporal scope of a theory). Whether we want a permissive or a restricted sense will probably depend on why we're asking the questions we're asking about feasibility. Is it for intellectual interest, or because we want to actually bring some outcome about?

These two points reveal something fairly important about the nature of feasibility assessments. They explains how easy it is for theorists using the concept to talk past each other. Suppose some theorist puts on the table the claim that a cosmopolitan global society is morally obligatory, given as it follows from many theories about fairness and equality between persons. Another theorist might claim that the outcome, a cosmopolitan global society, is infeasible. One might think it is infeasible because there is no agent *now* who can realize the outcome, while another might think it is feasible because there will be (or could be) an agent *at some point* who could realize the outcome. One might think it is feasible because there's a gerrymandered agent who could realize the outcome, while another might think it's infeasible because there's no 'genuine' agent who could realize the outcome (the line between 'genuine' and 'gerrymandered' agents

will always be fuzzy, but there are clear cases either side). These theorists are not really disagreeing; they both agree that it's infeasible to realize the outcome now, because no agent with the relevant option exists. And they might even agree that it could be feasible to realize it at some point in the future, either because agents now have an option to create an agent who could realize it, or because there is some gerrymandered agent who could realize it. What they would more likely disagree about is which of these is the *relevant* sense of feasibility for political theory. Should we care about feasibility now, or feasibility sometime? Should we restrict our attention to 'genuine' agents, or should we allow gerrymandered agents? The strongest sense is 'now', with 'genuine' agents, but there will be occasions where it is appropriate to deploy the weaker sense.

A final concern with the formulation in (e) is whether feasibility is something existential or something agent-relative. Consider the difference between 'it is feasible that Jens-Christian run 10 kilometres' and 'it is feasible for Jens-Christian to run 10 kilometres'. The grammar is only subtly different, but the success conditions are worlds apart. On the analysis in (e), the former is true so long as there is *some* agent who has an action available to him that could bring about the outcome that Jens-Christian runs 10 kilometres. Suppose that Weng Hong has the action available to him of holding hostage Jens-Christian's laptop in order to coerce him into running the 10 kilometres. In that case, the outcome that Jens-Christian runs 10 kilometres is feasible because Weng Hong has an action available to him that could bring it about. But that is certainly a bit strange. Shouldn't we rather care about what actions are available to *Jens-Christian*, and whether one of *them* could bring about the outcome that he run 10 kilometres?

The difference is interesting, and it points to the different roles feasibility can play mentioned in Chapter 1, and the way in which feasibility comes apart from 'ought implies can', discussed earlier in this chapter. 'Ought implies can' is concerned with individuals and their actions, while feasibility also extends to collectives and outcomes. The 'feasible-that' formulation is about outcomes, and it doesn't necessarily matter who brings them about, while the 'feasible-for' formulation is about the actions and outcomes available to a specific agent. In assigning blameworthiness we will care about 'feasible-for', because blame attaches to specific agents. Which of these two we are most interested in depends on the kind of theory we are assessing. If we take a broad theory about global institutional reforms, we might only care to ask whether there is *some* agent with an action that could bring about the theory's recommendations. But if we take for instance a consequentialist theory bearing directly on some agent, then what we'll be

interested in is whether *that agent* can do what the theory requires him to do. Imagine that a doctor desperately needs light so that he can perform an operation, and the only way to get light is if someone runs 10 kilometres on a kinetic energy machine. Then all we'll be interested in is whether there's some agent who can bring about the outcome of having run 10 kilometres on the machine. That agent might be Jens-Christian, but it might as well be someone else. Then the detached feasibility assessment in (e) is what we'd need. But imagine that some moral theory tells Jens-Christian that he has to run 10 kilometres, perhaps because only by running would he get to a location in time to save a damsel in distress. Then what matters is what's feasible *for Jens-Christian*. In that case, we had better supplement (e) with (f) to allow for both of these kinds of interests.

(e) The recommendations of some theory are feasible *iff* there exists an agent, or set of agents, within the designated period of time, who has available to her some action or set of actions that could bring the recommendations about.

(f) The recommendations of some theory are feasible for an agent *iff* she has, within the designated period of time, an action or set of actions in her set of options that could bring the recommendations about.

Both include three main variables: the agent, the time, and the action. Let's return to the Carrot Mob project introduced at the start of this section to illustrate replacing these variables. We can consider the proposition 'it is obligatory that cafes in German reduce their energy consumption'. The obligation is plausible *a fortiori*, because given climate change it is obligatory that *everyone* reduce their energy consumption. Now we can do one of two things. We can either think of all the actions that might realize this outcome, of which switching to more energy efficient appliances is one, and then think about whether there exists any agent who has one of those actions as an option, or who has as an option bringing into existence an agent who would have one of those actions as an option. Or we can do that the other way round, first thinking about which agents exist, and then thinking about what options they have, and whether one of those options is an action that would bring about the outcome of a cafe reducing its energy consumption. My proposal is that if there is some agent who has an action that would realize the outcome, *regardless of how likely the agent is to actually choose that action*, then the outcome is feasible. Consider the following:

(P1) It is feasible for *Krawall & Remidemi* to change their appliances to more energy efficient ones.
(P2) It is feasible for Carrot Mob to get *Krawall & Remidemi* to change their appliances to more energy efficient ones.
(P3) it is feasible for Clas to respond to Carrot Mob's campaign by going to *Krawall & Remidemi* on the required day.

In (P1) the agent is *Krawall & Remidemi*, in (P2) it is Carrot Mob, and in (P3) it is Clas. Who the agent is is important because we cannot assess the feasibility of a theory's recommendations without knowing *whose* actions would be required for the proposal to succeed, and whether that agent has the action as an option. (P1) is a claim about what the cafe can do. (P2) is a claim about what Carrot Mob can do. And (P3) is a claim about what Clas can do. If *Krawall & Remidemi* has as one of its options changing its appliances to more energy efficient ones, then (P1) is true. If Carrot Mob has as one of its options getting *Krawall & Remidemi* to chance its appliances, then (P2) is true, and if Clas has as one of his options going to *Krawall & Remidemi* on the required day to support the Carrot Mob campaign, then (P3) is true. The outcome of that cafe in particular reducing its energy consumption is feasible *because* there are agents who have as options bringing that outcome about.

So much for the structure of claims about feasibility, and what is important for determining whether they are true. Two crucial and related questions remain. The first is: when does an agent have some action as an option? If we don't know when some action is an option of an agent's, then this formulation of feasibility can do no work. The second is: how should we understand actions, to make room for getting others to do things? Normally an action is something that is under the full control of an agent. But seldom are the actions of others fully under my control, so surely including the option of getting someone else to do something as an action that could realize a certain outcome is not a good move in formulating feasibility in its binary, ruling out role. For what remains of this chapter I will focus on answering those two questions. In Section 4.3 I give the conditions for an agent's having an option, and in Section 4.4 I return to the question of actions and control over others.

4.3.1 Stability and Accessibility

To say that an agent has as one of her options an action that would realize a given outcome, certain facts must obtain. First of all, the agent must stand in a certain relation to that action. The action must be one that she can actually undertake. Second of all, the action must stand in a certain relation to the outcome. It must be one that could actually produce that outcome. Picking all the apples off the tree is not an action in Leon's option set, because he is in a wheelchair. And going to Fellows Garden for a beer instead of to the orchard to pick the apples is not an action that could produce the outcome that all the apples are picked from the tree. Drinking beer at Fellows instead is just not the kind of action that produces the outcome of apples being picked. We can capture this idea by way of the notion of *accessibility*. An outcome is accessible for an agent if there is a way that she can bring it about, a way for her to get 'there' from 'here'. In Section 3.3.1 I introduced Dennett's (1995) and Hawthorn's (1991) ideas of developmental paths being open at one time and forever lost at another. Accessibility can be thought about in this way, via the metaphor of a series of steps. Some outcome is accessible to an agent if there is a series of steps she can take in order to bring it about. An *action* is accessible to her if she can actually choose to do it; an *outcome* is accessible to her if one of her actions (or a set of her actions) could produce it.

Notice that I only say that the agent *could* choose the action, and the action *could* produce the outcome. This is a fairly weak notion of feasibility, as indeed is suitable for this fairly blunt binary role, in which feasibility is used to rule out the recommendations of theories which absolutely cannot be implemented. An action is not definitely ruled out so long as the agent could in principle choose to do the action in question (even if she almost certainly will not), and so long as that action could in principle realize the outcome (even if it almost certainly will not). If it is *certain* that going to Fellows will not lead to any apples being picked, then that action is not a feasible means to the outcome of all the apples being picked. And if going to Fellows is the only option available to Leon, then it is not feasible for him to pick all the apples, and thus the allegation that he is under an obligation to do so is false. But if going to Fellows has a 99% chance of producing the outcome that Leon is drunk and does nothing but flirt with pretty visiting students, and a 1% chance that he obtain some mechanical apple-picking equipment, then we should say that bringing about the outcome that all the apples are picked is feasible for him. *It is not ruled out.* There is an action in his option set, namely going to

Fellows, that has some chance of bringing about the outcome that the apples are picked. Thus in those circumstances, the allegation that he is under an obligation to realize the outcome of the apples being picked may well be true (I say may well be because whether a person has an obligation depends on antecedent moral commitments, and not just whether the person is able to fulfill that obligation).

What if there is an action that an agent could take to realize an outcome, but the 'realizing' would be in name alone? For example, imagine that some egalitarian moral theory requires a radical global redistribution of wealth. And imagine that there is some collective agent which has in its option set the action of appropriating huge amounts of wealth from developed countries under threat of nuclear attack. But imagine further that were the collective agent to undertake that action, developed countries would be so incensed by the method of appropriation that they would in all likelihood retaliate, leaving the prospects for a global egalitarian society rather slim. This is to suggest that there are some actions that would realize certain outcomes, *but only fleetingly*. Thus it might be a good idea to insist on a *stability* condition in addition to the *accessibility* condition discussed so far. An action is only a means to an outcome if it would produce that outcome in a way that would be more or less stable. Geoffrey Brennan and Philip Pettit stress the importance of stability considerations in their (2005), where they argue that political philosophers should be more concerned with 'the problem of how to ensure that whatever arrangements are put in place ... are arrangements that ordinary humans are able to sustain' (Brennan & Pettit, 2005, p. 264).[28]

We still need to know what *makes it the case* that an action is accessible to an agent, that a stable outcome is accessible by way of an action. What kinds of facts should we appeal to, to settle these matters? When a theorist makes a claim that the recommendations of his theory are feasible, how do we know whether his claim is true or not? Just as in the discussion of the 'can' part of 'ought implies can', we need to fix the facts that are the determinants of feasibility. There are some facts that make it the case that an action is *not* in an agent's option set, some facts that make it the case that an action will *not* produce a particular outcome, and some facts that make it the case that if an action does happen to produce a particular outcome, that outcome won't be stable.

28 Some might prefer to leave stability out of a conceptual account of feasibility, and in that case it could be repackaged as part of the desirability of some action. It is undesirable to put our efforts into pursuing an outcome that won't last (unless it is a transitional outcome). But it seems to me that the recommendations of political theories come with an implicit or explicit temporal scope, so that when we ask whether an outcome is feasible, what we're really asking is whether an outcome, for a given length of time, is feasible. The answer will be 'yes' if the outcome, for the given length of time, meets the conditions for binary feasibility.

What are these facts?

4.3.2 Feasibility as Possibility?

Although they will certainly not be the end of the story, facts about *what is possible* (in the philosopher's sense, rather than the ordinary language sense) certainly affect which actions are in an agent's option set, which actions can produce which outcomes, and which outcomes will be stable.

The standard way to model possibility is with a set of nested spheres (e.g. Dennett, 1995, p. 107), with the broadest kinds, like logical possibility, at the outermost layer, and with progressively narrower kinds toward the centre. Exactly what the layers are, and how they are ordered, differs from model to model. For example, there is a longstanding debate between Kripkeans and their opponents over whether metaphysical possibility, logical possibility, and conceivability are the same layer or different. If one thinks that ideal conceivability and logical possibility are the same, and denies Kripke's (1981) argument that discovered identities are metaphysically necessary yet logically distinct (e.g. Chalmers, 2002), then one would model these together as the outermost layer. If, however, one took a Kripkean line, one would nest metaphysical possibility *inside* logical possibility, taking the latter to be a broader category. Probably no political theory we encounter will make the mistake of violating metaphysical or logical possibility, or conceivability. So we need not stake a claim in that debate.

Nested inside the kinds of possibility mentioned already is nomological possibility. This means something like 'possible according to the laws of nature', the laws provided to us by our best sciences. If a political theory requires an outcome in which we travel faster than the speed of light (perhaps so that we can deliver the leftovers of our meals to the world's starving before they go cold?) then we can reject the requirement on the grounds that it is nomologically impossible. Or if a theory requires that we get rid of our rubbish by shooting it up into the sky and leaving it there stationary, we can say that while there is some action that would achieve that outcome, the outcome wouldn't be stable: the laws of nature provide that gravity would see it come right back down to earth (or to the nearest heavy object). But as with the kinds of possibility mentioned already, there are unlikely to be many political proposals so easy to rule out. We should take these kinds of possibility to provide the kinds of facts that would rule out an action for an agent, rule out an action's producing an outcome, and

rule out an option's being stable. But if we want the account to do any real work, it will have to rule out more than that.

Both Richard Dawkins (1986) and Daniel Dennett (1995, esp. Ch. 5; 1986, esp. Ch. 17) offer accounts of *biological* possibility, which stand to do quite a bit more ruling out. Dennett uses the metaphor of a library containing all possible books, to anchor his thought-experiment of a library containing the descriptions of all possible genomes. The lesson he wants to draw from the metaphor is that it is not the case that for any description of a possible genome, that genome would be viable. We can crudely conceive of such descriptions as recipes. The philosophical point is that while all such recipes might specify 'something', they do not specify a viable organism. Most of the recipes are gibberish (Dennett, 1995, p. 113). Such a thought-experiment highlights the real distance between a logically possible genome, and a biologically possible one.

Dennett points out several complications which serve to further narrow the space of what is biologically possible. One is that biological possibility must always be tacitly affixed to a given time, because viability is relative to environment (he describes it as 'a moving target') (Dennett, 1995, p. 116). He argues that we must specify a starting point when discussing restricted notions of possibility; we must ask not 'what is possible', but 'what is possible *now*?' (or at, or from, some specified point *p* at which we hold fixed environment, state of knowledge, and background conditions), and we must specify the travel parameters, i.e. what things are to be allowed when we think about getting there from here (or here from there, if our target is historical) (Dennett, 1995, p. 118). Absent certain conditions, a given organism is not viable. Whether an organism is viable depends on how suitable the environment is, whether an appropriate 'host' exists, and whether manipulations of the genotype are able to produce the desired changes in the phenotype. Tiny changes in a genotype can produce very large changes in the phenotype. The implication is that some intermediary step between two manifest phenotypes might be biologically impossible, because the distance between them was specified by the smallest possible permutation of the genotype. So it seems that if we take all these things into account, we have a much narrower pool of biologically possible organisms than were conceivable under the genome-library thought-experiment, and under a less circumscribed kind of possibility. Feasibility assessments apply to political theories and prescriptions, which have as their subject matter outcomes that human action and interaction might produce. Presumably biological possibility takes account of all living organisms. So insofar as this kind of possibility provides the right kinds of facts,

we might want to limit our attention to *human* biological possibility. Again though, not many political theories are about the biological adaptations we ought to evolve. Some political theories are, however, about *cultural* adaptations we ought to evolve.

Biological theorists are divided over how seriously they take cultural influences. Some refuse to take it seriously at all, and talk solely in terms of genotypes and their heritability. Others integrate genotypic influences with *phenotypic* influences, taking the effects of cultural inheritance to be a defining feature of human development (e.g. Boyd and Richerson 2005; Bowles and Gintis, 2003). This school of thought is known as gene-culture co-evolution. It is easy to imagine a political theory proposing reforms that clash with biology or culture (especially given the wide range of behaviours specific to different cultures around the world). For example, a theory aimed at preventing suffering in a time of increased pressure on resources might recommend that we stop having children (at least for a while), which would be all but biologically impossible, or that we start caring about distant strangers as much as we care about kin – in fact, several theories do propose this – which would be all but culturally impossible, at least in light of the biological story about the evolution of culture, in which preferences toward helping those near to us have been evolutionarily adaptive.

Taking facts about the kinds of organisms we are as relevant to making binary feasibility assessments makes us susceptible to controversial disputes over human nature. Is there any such thing? If there is, what does it include? Facts about what is biologically and culturally possible are something subtly different from facts about the essential nature (if there is such a thing) of persons. There are certain developmental paths that are not open to members of our species. The fact that these are not open will mean that certain kinds of proposed reforms are simply inaccessible to, infeasible for, people like us. That doesn't mean we should take human option sets to be limited by the generalizations that come along with the 'human nature' idea. These generalizations include that men are by nature unfaithful to their partners, that women are by nature suited to domestic duties, that people will in general seek status and esteem, that the stronger will seek to exploit the weaker, and so on. All of these claims and more have featured at some point in the contemporary media in the ongoing debate over whether there is such a thing as a human nature, whether any of our behavioural traits are innate or essential, and what the relative importance of genes versus environment is.

Evolutionary biologists have shown rather clearly that the idea that people's behaviour is genetically determined is highly improbable. As Alex Rosenberg and Daniel

McShea summarize it, (a) it is difficult to individuate genes according to their functions, firstly because the generation of a single specific protein or enzyme requires a large number of different and disparate nucleic acid sequences, secondly because the same functional gene is multiply-realizable on a variety of different sequences, and thirdly because any number of genetic and environmental starting points can create pathways to lead to the creation of the same enzyme; (b) very few of the studies showing a correlation between genes and behaviour (e.g. genes for violence, or alcoholism) have been replicated; and (c) genes are multiply-realizable depending on environmental stimulus, and a phenotype (observable characteristics of an individual, e.g. behaviour) *just is* the joint product of genes and environment, so obviously genes exposed to different environments will manifest different phenotypes (Rosenberg & McShea, 2008, pp. 208-209). Thus the fact that two individuals have the same genotype does not mean they will have the same phenotype. So instead of saying that there is 'a gene for *x*', it's more accurate to say something like that there is 'a gene for *x, y* or *z,* depending on the environment'.

The degree of phenotypic variation given a genotype, and the extent to which human behaviour is plastic, given our raw genetic material, are both hugely interesting questions. And at the very edges of inquiry about which outcomes it is possible for people to bring about, they will be relevant, and the fact that we do not know the answers to them will mean that we cannot say whether a certain political proposal is feasible or not. Perhaps a cosmopolitan theory will one day require that all human adults care for any human child made salient to him or her. If and when that theory is put forward, we will have to ask whether humans are developmentally plastic enough to throw off their long history of caring for their own kin (and investing in their own genes). But those kinds of cases are unlikely to be central. Thus I shall for the most part bracket the interesting questions about developmental plasticity when talking about the feasibility of political outcomes. Certainly I think culture can have a strong influence on people, but I do not think we should include cultural constraints as the kinds of facts that rule a proposal out of consideration for implementation. I will argue in Chapter 5 that we should include cultural constraints as soft constraints for the graded sense of feasibility to be discussed there, which affects how feasible a proposal is, but not whether it is feasible.

When we ask questions about whether certain outcomes can be brought about, the answers are limited by the more and less restricted kinds of possibility discussed so

far. But at heart, the questions have to do with the abilities and capacities of ordinary agents, both individual and collective. We want to know whether there is some agent who has the ability or the capacity to bring about the allegedly obligatory outcome. Perhaps whether some agent has the ability or the capacity will depend on what institutions and technologies the outcome depends on, and whether they exist, or can be brought to exist, within the designated time frame. Whether particular technologies and institutions exist or can be brought to exist depends in turn on the abilities and capacities of human agents, both as individuals and as members of groups. In the next section I shall concentrate on proposing a formulation of feasibility in terms of 'hard constraints'.

4.3.3 Hard constraints

Logical and metaphysical possibility, conceivability and nomological possibility are all part of the 'hard constraints' upon a political theory or recommendation, which is to say, if instantiating that theory, or realizing those recommendations, would be impossible in light of those facts about what is possible, then the theory or the recommendations can be rejected on the grounds that it is infeasible. It is infeasible in the binary sense of being ruled out. Violating a hard constraint is proof that a theory cannot be realized in practice. The kinds of constraints mentioned so far I would refer to as *timeless* hard constraints. They will for all intents and purposes constrain what is feasible for all of our time on this planet.

But there are other hard constraints, violation of which makes it impossible to instantiate some theory or realize some recommendation, but which are *not* necessarily constraints for all of our time on this planet. The biological constraints mentioned earlier are plausibly in that category. Perhaps at some point in the future, biotechnology will be such that we can push the limits of what is biologically possible, through genetic enhancement, genetic engineering, and other technologies. (And as I have suggested, the cultural constraints already mentioned are not hard constraints at all. We have different kinds of social practices and normative commitments depending on the culture we are a part of, but the kinds of practices and commitments we have are more or less malleable. It might be very difficult to change them, but it is not impossible. I return to these kinds of constraints in Chapter 5).

Biological constraints are part of a sub-category of hard constraints that are time-indexed. That is to say, they are properly regarded as hard constraints, because they

make it impossible to implement some theory or recommendation. But they are not *timeless* hard constraints, constraints that make a proposal impossible to ever implement, as some of the constraints already discussed are. They are time-indexed because they act as rulers-out now, but not necessarily in the future. Other kinds of time-indexed hard constraints include technology and medicine, and in some cases resource and financial constraints. We can work at the frontiers of medical and technological progress, but we can go no further than that. The lack of an existing technology now makes it impossible to implement a proposal that requires it, but once the technology has been developed, that will be no constraint at all. There are machines we cannot yet build, and medical operations we cannot yet perform, and these are hard constraints on what we can do now, but not on what we can ever do.

The same story goes for resource and finance constraints. If a country has access, at best, to only so much oil, then that amount of oil is a hard constraint on its oil-related plans. Or if it can borrow, at best, only so much money, then that budgetary constraint is a hard constraint on its domestic and international plans. We might also want to include context-dependent constraints as time-indexed hard constraints. I cannot shoot the hostage-taker and rescue the hostages because I do not have a gun. It might be possible that I go and get a gun, i.e. there is something I can do to make it the case that I can shoot the hostage-taker and rescue the hostages, but the fact that I do not have access to a gun *now* is a hard constraint on what I can do *now*. Likewise if I do not speak German, I cannot converse in German with my German friends now, even though there's something I can do, namely take German classes, so that I can converse with them in German later on. Context-relative constraints are time-indexed hard constraints, but they are not timeless because we can do something to disable them. The violation of timeless hard constraints suffices to rule a theory or recommendation out as permanently infeasible; the violation of time-indexed hard constraints suffices more weakly to rule a theory or recommendation out as infeasible for the duration that the context remains such as to sustain the inability or incapacity.

4.4 A Binary Feasibility Test

I have argued that in its binary role, feasibility constraints follow the structure of 'ought implies can' constraints, in that they are primarily concerned with ruling out theories requiring outcomes that cannot be implemented, or recommendations that

cannot be realized, in the same way that 'ought implies can' is concerned with ruling out obligations that people cannot possibly fulfill. A recommendation is feasible in this binary sense if there is some agent who has an action in her option set that could realize it. Actions are in option sets when they are *accessible* to agents. Outcomes are accessible to agents when an action in their option set has some chance of bringing them about. And finally, a *stable* outcome is accessible to an agent when an action in her option set has some chance of bringing it about non-fleetingly. Which actions and outcomes are accessible, and which outcomes are stable given actions, is *determined* by whether the agent's doing that action, the action's producing that outcome, or the action's producing a *stable* outcome, are ruled out by timeless hard-constraints, or by time-indexed hard constraints for the entire duration of the temporal scope of the recommendations. If a theory recommends that I speak German *right now*, and I do not know how to speak German, then the fact that the recommendation violates this time-indexed hard constraint on the actions which are in my option set allows me to reject the theory or recommendation. If a theory simply recommends that I speak German, then the recommendation is feasible in the binary sense so long as there's something I can do to make it the case that I can realize it. There is something I can do; I can learn German. So the recommendation is feasible. Thus from (e) and (f), the following tests of binary feasibility:

(General) Binary Feasibility Test: The recommendations of some theory are feasible *iff* there exists an agent, or set of agents, within the designated period of time, who has available to her some action or set of actions that could bring the recommendations about.

(Agent-relative) Binary Feasibility Test: The recommendations of some theory are feasible for an agent *iff* she has, within the designated period of time, an action or set of actions in her set of options that could bring the recommendations about.

Truth conditions: An agent has available to her some action *iff* her performing that action is not ruled out by any hard constraint. An action could bring the recommendations about *iff* the action's producing the outcome is not ruled out by any hard constraint (an action could bring the recommendations about stably *iff*

the action's producing the outcome stably is not ruled out by any hard constraint).

There are some problems with this proposal. One is the second crucial question mentioned at the end of Section 4.3, namely, a problem with the meaning of 'agent'. The other is to do with saying when an action 'could' bring an outcome about. Let's take these in order. We establish feasibility by looking at what is feasible *for* any given agent. But what counts as an agent? Both individual and collective agents must count, because often the outcomes we will be asking about will be such that only collective agents can bring them about. When we want to know, for example, whether it is feasible that the New Zealand government provide redress to Maori for historical injustices, we want to know which actions are available (i.e. in the option set of) *New Zealand*. But how about future agents? And how about gerrymandered agents? What if there is no 'genuine' agent who has an option that could realize some recommendation, but there is a gerrymandered agent who could? What if no current agent has an option that could realize some recommendation, but current agents have the option of *creating* a collective agent which *could* realize the recommendation?

First of all, neither future agents nor gerrymandered agents are agents of the right kind to feature in the agent-relative binary feasibility test. They are not 'genuine' agents. I use scare quotes to indicate that the line between genuine and non-genuine agents is a difficult one. Certainly there are clear cases of individual agents, and clear cases of collective agents. It is easy enough to rule future agents out with this line: agents are things that currently exist, or that exist within the temporal scope of the theory. For the general test, I said in Section 4.3 that future agents count so long as they could be brought to exist within the temporal scope of the recommendations, although sometimes we will restrict our focus to currently existing agents only. If a theory recommends some outcome be brought about *now* we ask whether some agent exists *now* who has available to her an action that could bring about that outcome. If a theory recommends some outcome be brought about within a definite period of time, we ask whether some agent will, or could be brought to, exist within the given period of time with an action available to her that could bring about that outcome. The trickiest cases are recommendations with open-ended temporal scope, such as 'it is obligatory to reduce suffering'. Presumably this obligation ends when there is no more suffering. Thus, so long as there is suffering (and it is feasible to reduce it), the obligation stands. These recommendations are tricky because, so long as there will be *some* agent *sometime* who can

reduce suffering, they are feasible. But that seems like the right answer. We wouldn't want to rule reducing suffering out as infeasible just because it happened to be the case that no current agent could do anything to reduce it.

It is less easy to deal with gerrymandered collective agents. Why is New Zealand a collective agent, and the 'global society' not a collective agent? Giving a precise account of agency will be similar to arguing against arbitrary mereological fusions in the ontological debate. Certainly a pen is an object, and my left finger, your right eyeball, that kitten over there, and a faucet in China probably is not an object. But is the kitchen one object, or a set of objects? Is a table one object, or are the legs and tabletop an object each? Philosophers (at least in Australasia) tend to allow arbitrary fusions into their ontology, and say simply that some objects are more *useful* than others. I could make a similar concession for collectives, and say that all kinds of collective agents, even the most gerrymandered, are agents, but some agents are more useful than others, namely, the ones we might actually expect to *do* something. On the other hand, I can avoid that concession, by saying that collective agents are groups with the right kind of internal structure. Then the details come in saying what the right kind of internal structure is. I talk about these details in Chapter 8, following Alexander Wendt and Peter French.

So much for the first problem, saying what is within the scope of 'agent'. The second problem had to do with saying when an action 'could' bring an outcome about. I said earlier that Leon's going to Fellows Garden for a beer is 'not the kind of thing' that causes apples to be picked from trees in an orchard. His going to the orchard and picking the apples from the trees is the kind of thing that causes them to be picked. When I drop a porcelain plate on a concrete floor, the laws of nature make it such that the plate smashes. So when I ask whether the outcome of the plate not smashing is accessible via the action of dropping it on the floor, I would say that it is not: the action is an infeasible means to the outcome. But now consider strange events with an extremely low quantum mechanical or statistical mechanical chance. These are vanishingly rare; almost certainly no one of us will ever witness one in our lifetime. But they are possible, with some tiny probability. If such an event were to occur, the plate would not break, but let's say would rather fly sideways and land safely on a pile of cushions (example from Hawthorne, 2005, p. 396). So, with the possibility of these kinds of events in mind, none of the hard constraints I mentioned bearing on an action's producing an outcome actually rule out anything. There's always *some* tiny chance that

an event like this occurs and brings some totally unexpected outcome about.

What this means is that the part of the binary feasibility test that says 'an action could bring the recommendations about *iff* the action's producing the outcome is not ruled out by any hard constraint' cannot be quite right. The hard constraints do not rule out anything if they allow for these strange events with extremely low quantum- or statistical mechanical chance, and facts about what is conceptually, metaphysically, nomologically and biologically possible all allow for those strange events. The problem is that we cannot just add a stipulation to the binary test along the lines 'barring the possibility of an extraordinarily unlikely event'. That is because if we deliberately set up many upon many (*ad infinitum*) instances of the same action, we would *expect* to see that action produce an outcome that was extremely unlikely. And it would be strange to have discounted that possibility by stipulation. An action is a feasible means to an outcome if there's some chance it will bring it about, and this extends to extremely low chances. But I can nonetheless say for practical purposes that events with a radically low chance, such as these kinds of events have, can be taken as ruled out. It is not strictly true that they are, but it is okay to work on the assumption that they are.[29]

4.5 'Abilities': a parallel discussion

I said in Chapter 1 that another role for feasibility is to tell us what abilities (or capacities, or powers) things have. What I will say in this section is meant to contribute to understanding feasibility in that role. The philosophical discussion of 'abilities' follows on from the influential dispositional account of abilities defended by Gilbert Ryle (1949). The account of abilities has been more focused upon individuals and their actions, but the question of what abilities collectives have is interesting, and under-explored (more on collective abilities in Chapter 7). We can easily reformulate the binary feasibility test in terms of abilities, for example: 'the recommendations of some theory are feasible *iff* there exists an agent ... who has the *ability* to bring the recommendations about'. The most prominent analysis of what it is to have an ability is the conditional analysis:

29 It's difficult to draw a convincing line between events with 'radically' low chances of happening and events with slightly higher chances. Given this, it might be better in the end to deny that anything is really 'ruled out', and say that all senses of feasibility are, at base, graded. Although that bears thinking about, it is a more controversial thesis than I am prepared to defend here.

Conditional Analysis: *S* has the ability to *A* iff *S* would *A* if *S* tried to *A* (Maier, 2010).

In the conditional analysis, an agent has an ability if she *would* act if she *tried to* act. So, Jens-Christian has the ability to run 10 kilometres if he would run 10 kilometres if he tried to. In the general binary test, the outcome of Jens-Christian running 10 kilometres is feasible if there is some agent who has an action available to him that would bring about Jens-Christian running 10 kilometres. But notice how different these two conceptions are. Abilities are indexed to agents. When we ask whether an agent has an ability, we look at what *that* agent would do if he tried (assuming for now that the conditional analysis is right). But with the general binary test, when we ask whether an outcome can be brought about it only matters whether there is *some* agent that could bring it about. For example, if Weng Hong could make it the case that Jens-Christian ran 10 kilometres, let's say by bribing or coercing him, then the outcome of Jens-Christian's having run 10 kilometres is feasible. We might say, however, that which agents are relevant to a feasibility assessment are established by context, which is to say, what a theory recommends tells us whether the general or the agent-relative binary test is appropriate. You might think that when we are asking whether it's feasible that Jens-Christian run 10 kilometres, we are really asking about what is feasible *for* Jens-Christian, not about what is existentially feasible. And when we are concerned with this agent-relative sense of feasibility, then we are asking much the same question as whether an agent has the *ability* to act in a way that has a chance of producing the desired outcome. There are two well-known objections to the conditional analysis of abilities, however. If the agent-relative binary feasibility test is kin with the discussion of abilities, then we might worry that these objections to the conditional analysis of abilities apply as well to it.

The first objection is to the sufficiency of the analysis, raising counterexamples in which *S* cannot try. For example, if determinism is true, then no one is able to do anything other than what they actually do. Or in another example, a person may be psychologically unable to (bring herself to) do something, even though able in the normal sense. The second is an objection to the necessity of the analysis. Maier discusses the example of a very good golfer missing an easy putt. In that case we shouldn't think it's true that he would make the putt if he tried to, because he did try, and he failed (Maier, 2010).

We can surely set the issue of determinism aside here, firstly because, as Maier points out, the conditional analysis might be true in a deterministic world (Maier, 2010). The determinist constraint would simply entail that people have the ability to do a lot of things that they don't end up doing. But it is plausible that people have abilities that they do not exercise, so this is not a terrible blow to the analysis. And secondly, we can set it aside because in normative philosophy we generally proceed on the assumption that people are able to do other than they actually do. The counterexample in psychological terms is more interesting, but I can't help but think it relies on an equivocation between different sense of ability. Either the psychological sense is part of ability, or it is not. If it is, then we shouldn't assume it as part of 'trying'. Just as we would say that *S* does not have the ability to play the piano if it came out false that she would if she tried to, we should say that she does not have the ability to pat a spider if it comes out false that she would if she tried to. The fact that it is a psychological mechanism that would cause her to fail shouldn't bear on our judging it to be a genuine failure. The instinct to treat the psychological case as a counterexample to the conditional analysis of abilities probably comes from thinking of psychological constraints as things that will stop a person from trying. If that were the correct analysis, then the counterexample would go through (*if* she were to try to pet the spider she would actually do it, but she'll never try, because she's afraid of spiders). But I see no reason why we should treat them in this way. It's just as plausible to think that if she tries she'll fail, because her fear will get the better of her.

The objection to the necessity of the conditional analysis seems also to equivocate between different senses of ability. In one sense, the golfer has the ability to make the easy putt; he's a good golfer after all, and good golfers are in general able to make easy putts. But in another sense he doesn't, because he didn't. Maier suggests we might get around this using a distinction between general and specific abilities, saying that the golfer has the general ability to make the putt (he has the necessary skills) but not the specific ability (he can't do it under those particular conditions) (Maier, 2010). Other objections to the necessity of the account come from the discussion over dispositions and include cases of finking, masking, mimicking, and antidotes (Fara, 2006). These are designed to show that sometimes a person does look to have an ability, but when they try to do something they fail for unusual reasons.

For example, the Frankfurt cases prominent in the free will debate are cases of finking (Frankfurt, 1969). A person might have prima facie an ability to save a drowning child, but unbeknownst to him if he chooses to save the child a malevolent psychologist

will interfere with his brain signals to make it the case that he chooses not to save the child. The malevolent psychologist acts as a fink. Or to give a case of masking, imagine we say that a vase is fragile because it would break if it were dropped. Wrapping it in styrofoam is a case of masking, because then we could drop it without it breaking. They styrofoam masks the vase's disposition to break. Antidote cases are such that we analyze poison as having the disposition to kill a person if consumed, but in which the poison can be consumed and yet fail to kill so long as the consumer take the relevant antidote. Finally, imagine that a sturdy metal cup is such that it would break if dropped, but only because a malevolent angel is waiting nearby and will make it the case that if the cup is dropped it will shatter. These are cases of mimicking, and they are the reverse of finking cases (on finking cases see Martin, 1994; on masking cases see Johnston, 1992; on mimicking cases see Smith 1977; Johnston, 1992; on antidote cases see Bird, 1998; and in general see discussion in Fara, 2006).

There are many proposed solutions to these various kinds of objections, some of which try to build a *ceteris paribus* clause into the account. For dispositions that would be to say that, for example, a glass vase is fragile if it would break when dropped *under normal conditions*. Or for abilities, that a person has an ability to swim only if he would swim if he tried *under normal conditions* to swim. At least in Maier's exposition, the biggest challenge to the conditional account of abilities are the masking cases. But these seem actually to be the least challenging. No one thinks that a vase is fragile only if it has a disposition to shatter when dropped. If it is dropped onto a bed of feathers, or a large pile of sand, or a heap of soft cushions, it most likely will not break. So it cannot be merely the dropping that is important. It is rather the conditions more generally. When the vase is dropped such that it makes contact with a hard surface, and there are no countervailing factors such as an intervenor who will catch the vase before it hits the ground, the vase will break. But it is easy to see that wrapping the vase in styrofoam is just like dropping it into a pile of feathers. This is just as we think a correct account of colour is how it appears in regular viewing conditions. If we add a strong red light while viewing a yellow cube, we don't thereby fail in saying that the cube has a disposition to look yellow. We can't say how it has a disposition to look, because we are not viewing it in normal conditions. Masking cases then don't seem like counterexamples at all; they merely alter the conditions such that we wouldn't predict the disposition to be realized.

Notice that objections from finking, masking, antidotes and mimicking are all an artifact of the analysis of abilities in terms of what *would* be the case. The conditional

analysis claims that an ability is present so long as a person would *A* if she tried to *A*. But in the account of feasibility I have defended in this chapter, it is not necessary that the *A*-ing actually obtains. Some action is feasible as long as it is in the agent's option set, so long as she has the ability to do it. Some outcome is feasible so long as an action in the agent's option set *could* bring it about. Cases where she *wouldn't* when she tried aren't enough to show that she *couldn't*. They impact the conditional analysis of abilities but not my picture of hard constraints ruling actions in or out.

The parallel between the account of feasibility and an account of abilities should be obvious. I will mention the further parallel to an account of dispositions, because some recent theorists in giving a diagnosis of abilities have argued for a "new dispositionalism", which deals with abilities in a dispositionalist framework. Michael Fara for example gives the following account:

> *S* has the ability to *A* in circumstances *C* iff she has the disposition to *A* when, in circumstances *C*, she tries to *A* (Fara, 2008, p. 848; cited in Maier, 2010).

Of course we cannot understand this proposal without the relevant account of dispositions which is embedded within it. Fara's own account of dispositions is as follows: "an object is disposed to *M* when *C* if [and only i]f it has an intrinsic property in virtue of which it *M*s when *C* (Fara, 2006; discussing Fara, 2005). So what is the account of abilities being proposed here? Let's take the case of being able to swim. A person is disposed to swim when in deep water only if she has an intrinsic property in virtue of which she swims when in deep water. And she has an *ability* to swim only if she has the disposition to swim when she is in deep water and she tries to. What is plausible about a dispositional account of abilities is that it treats an ability as a special property of a person, one that is not always manifest, and can in fact be present even if never manifest. A glass can have the disposition to shatter if dropped even when it is never actually dropped.

What is peculiar about an analysis in these terms is the way it has to bring in intrinsic properties and assume trying. A glass has a disposition to shatter when dropped. But it seems weird to replace a condition like 'when dropped' with a condition like 'when she tries'. After all, trying is not obviously a condition that obtains in the world; it is a mental state or an action, and thus plausibly the subject of exactly the analysis we are trying to give (not formally external to the object in the way being dropped is). It is an interesting question to what extent a person has an ability to try. Thus I should think it

would be preferable to try to work with some version of the conditional analysis of abilities, rather than moving to this kind of analysis in terms of dispositions.[30]

Despite the similarities, there are also many differences between the account of abilities and an account of comparative feasibility. I have already said that abilities deal with individuals and actions. Feasibility deals with those things too, but also with collectives, and with states of affairs (outcomes). Also ability on the most prominent conditional analysis is about what *would* happen given trying. Trying is usually spelled out intentionally or volitionally. But on my account, it is a matter of what could happen given trying. And trying is itself an action, so if there are further questions to be asked about the feasibility of trying (ability to try) I have not asked them. I would deny that the counterexamples to the conditional analysis of ability, although structurally similar enough to warrant attention, succeed in undermining the binary feasibility test presented in this chapter. The argument from necessity seems to equivocate between various senses of ability, and the argument from sufficiency does the same, and when it does not falls back on old-fashioned worries about a deterministic universe, which we can simply assume away from the purpose of a project like this one.

A theory of abilities is desirable because of the way the concept of abilities has been used, unexplained, in philosophical accounts of concepts, of knowledge, and of "knowing what it's like" (see discussion in Maier, 2010, Section 5). But a theory of feasibility is desirable in helping non-ideal theorists to do non-ideal theory, in playing an important role in restricting normative political prescriptions. You might think the different objectives of the two concepts are sufficient to warrant a quite different treatment. Finally, the problem with counterexamples in the case of abilities and capacities is that they clash with our intuitive understanding. But there is no understanding of feasibility that is both intuitive and reliable – that is part of the reason a project like this one is needed. In a way this means we have more freedom, because the theory determines what is and isn't feasible. Unless the result that some action is infeasible really clashes with commonsense, the fact that a good theory suggests it is a reason to believe it, not a reason to revise the theory.

If that is right, then what I have said in the chapter prior to this section suffices to

30 There are two broad strategies of giving diagnoses of abilities, one conditional, and the other in terms of restricted possibility (see e.g. Maier, 2010). These correspond methodologically to the binary and comparative accounts of feasibility I have defended here. The binary sense is a restricted version of possibility (something is feasible if it does not violate certain relevant kinds of possibility), and the comparative sense is conditional (upon likelihood of success, given trying). I discovered this structural similarity only after having defended the binary and comparative accounts, and too late to make more extensive use of it.

an account of ability. That is to say, an agent has an ability when she has an action in her option set. What is ruled in and what is ruled out of her option set depends on the hard constraints upon her action. Both existential constraints (the various kinds of possibility) and agent-relative, context-relative constraints (the things she cannot do because she lacks the opportunity, or the skills, even though she might come to have them in another context) do some ruling out. An agent does not have the ability to perform an action if it is not in her option set, and she does not have the ability to bring about an outcome if the outcome is inaccessible by way of any of the actions in her option set.

The binary feasibility test is *not* a conditional analysis like the most prominent analysis of abilities is. The binary test uses a conditional probability. An action is accessible to an agent when it has a non-zero probability of being performed, *given* the agent's choice, or intention, or trying, to perform it. And an outcome is accessible by way of an action when the outcome has a non-zero probability of coming about, *given* the action. And a conditional probability, e.g. the probability of X given Y, is different to a conditional, e.g. the probability that if X then Y.

4.6 Conclusion

In this chapter I have tried to develop a binary test of feasibility, following the structure of the principle that 'ought implies can'. I said that while recommendations might be made in isolation from agents, we must assess their feasibility by asking whether there is any agent who has an action that could realize those recommendations available to her. These agents can be individual or collective. If they are collective, they must be of the right kind, i.e. with the right kind of internal structure (more on this in Chapter 8). And the agents must exist within the temporal scope of the recommendations. Whether an agent has some action available to her or not, and an action is of a kind that could realize a theory's recommendations, is established by whether the agent's acting or the action's producing the given outcome violate hard constraints, both timeless and time-indexed sensitive to the temporal scope of the recommendation, and for practical purposes barring the possibility of strange events with extraordinarily low chance.

I have also considered objections to the conditional analysis of abilities as objections to the binary tests, but denied that they are worrying, and suggested rather

than the binary tests provide an account of abilities, which contributes to feasibility's role of telling us what powers things have.

Notice that in this particular role, the feasibility test establishes what is *minimally* feasible, that is, what recommendations we cannot rule out as unrealizable. While this is part of the story about feasibility, it is not the whole story, as I suggested in Chapter 1 when I mentioned the various different roles that the concept of feasibility can play. In the next chapter I will try to develop feasibility in another role, namely in the graded role that allows for comparison between alternative theories. This is the role in which feasibility departs from accounts such as Brock's or Buchanan's (discussed in Chapter 3). It is not enough to know that there *is* some pathway from an action to a desired outcome. Rather we have to know *how probable* it is that the action will produce the outcome. This in the end may be the sense of feasibility that does the most work; although it will presumably be more controversial than feasibility in its binary role (people are used to thinking about 'ought implies can', after all).

Chapter 5

Political Feasibility II: The Comparative Sense

5.1 Introduction

In Chapter 4, I argued that if feasibility constraints in political theory were to follow the role of 'ought implies can' constraints in moral philosophy, then they would be limited to the ruling out role in which unrealizable proposals would cease to have any normative force. But I said in Chapter 1 that feasibility constraints have a further role to play. They must also allow comparative assessments of alternative political proposals. Having a good understanding of feasibility in this role will allow us to rank various proposals according to how feasible they are, which, when we introduce assessments of how desirable they are, will allow us to make the relevant tradeoffs between feasibility and desirability in making all things considered judgements about what we ought, politically speaking, to do (more on these judgements in Chapter 6). In this chapter I will be concerned exclusively with feasibility in this comparative role. I will argue that comparative feasibility uses conditional probability. What matters is the probability of an outcome *given* some action of an agent, where that action is feasible in the binary sense for the agent.

5.2 Structure of the concept

In the last chapter I gave the following as the basic structure of a claim about feasibility:

> **Binary Feasibility Test:** The recommendations of some theory are feasible *iff* there exists an agent, or set of agents, within the designated period of time, who has available to her some action or set of actions that could bring the recommendations about.

What matters to establishing whether a theory is feasible or not is whether there is some agent who has a relevant action in her option set. But this is the structure of feasibility in the binary role. The recommendations of a theory are either true, or they are not. What is the structure of a comparative claim? I suggest, preliminarily,

something like the following:

> (a) The degree to which the recommendations of a theory are feasible is established by....

The same variables will be in play, namely some agent and the set of options available to her. But I need to introduce a probabilistic element. In what follows, I shall concentrate on filling in the details necessary to complete the analysis of feasibility in its graded role.

5.3 Stable, accessible, and…? Likely to succeed

In the last chapter, I argued that feasibility in its binary role places hard constraints upon a theory. An outcome must be accessible, which is to say, there must be a path from 'here' to 'there', from our current context (or the desired point of departure) to the desired state of affairs, and it must be stable, which is to say, must be likely to endure once it has been brought about. If a theory requires an outcome that is both stable and accessible, and no hard constraints rule out the actions required to produce it, then the proposal is feasible in the binary sense. But as I have already mentioned, those constraints are not sufficient to the comparative role of feasibility. In that role, it is not enough that a pathway merely exist, that pathway has in addition to be *reasonably likely to lead from 'here' to 'there'*. Imagine that a theory requires an outcome for which there is a pathway, but along which the chance of our taking each of the steps is very low, and thus the chance of the outcome is very low indeed. We shouldn't say that outcome is more feasible than another with a higher chance of success. Thus probability plays a crucial role in the comparative sense of feasibility. But it cannot be *just* chance of success that matters: we mustn't confuse feasibility with likelihood. In the rest of this chapter I shall try to say how probability is relevant, suggesting that comparative feasibility is established by the extent to which an action producing a particular outcome clashes with 'soft constraints.'

Another caveat is necessary at this point, which is that it is important to distinguish the ontological claim about feasibility from the practical claim. For all I know, it might be that ontologically, every proposal is *either feasible or not*, in the binary sense. Then it would be the fact that we do not know this that makes the comparative test necessary, not that a proposal is genuinely feasible to degree *x*. The fact that we are

epistemically limited makes it the case that we can be wrong about the degree to which a proposal is feasible. We might judge on the basis of the available evidence that a proposal is highly likely to succeed, and turn out to have been wrong (more on this in Chapter 6).

5.4 Soft constraints of various kinds

Under what conditions is an outcome's being produced *more feasible*? Which is to say, how am I to fill in the details of (a) for the strongest possible graded analysis of feasibility? In Chapter 4 I argued that an outcome (i.e. some recommendation being realized) is feasible in the binary sense whenever there exists an agent who has an action available to her that could produce the outcome. I said that what determines whether she has such an action available to her is whether hard constraints rule out either her performing that action, or the action's producing the particular outcome. But hard constraints provide only a threshold. So long as a proposal does not violate hard constraints it is strictly feasible, but nothing in the hard constraints tells us whether one proposal is more feasible than another. That is why feasibility in this second graded role is important. We need to know not only what rules out an outcome, but also what makes it *more or less likely to come about*. What kinds of constraints are relevant here? I will refer to these as 'soft' constraints, in contrast to the 'hard' constraints of the last chapter. 'Soft' indicates that they are malleable, unlike their counterparts in the ruling out role. I will argue that the more an action producing an outcome is made unlikely by the relevant soft constraints, the less likely it is to succeed, and therefore the less feasible it is. Notice that at this stage there are no longer two steps, an agent's acting *and* the action's producing an outcome. Rather all that matters is the second step, and in this case it is the *extent to which* the action is likely to produce the outcome. The actions are defined relevant to the agent as those that are within his option set. Which actions are in an agent's option set is determined by the binary feasibility test presented in the last chapter. I will say more about why in Section 5.4.4.

5.4.1 Economic, institutional, cultural

The three most obvious kinds of soft constraints upon an action's bringing about some outcome are economic, institutional, and cultural. The first are the constraints

associated with the economic system, the second are the constraints associated with various political institutions, and the third are the constraints associated with culture – including religion – all of which characterize the context in which an agent's action takes place. None of these are hard constraints, because it is not impossible to violate them, but they are soft constraints (rather than not constraints at all) because their existence makes a proposal that clashes with them less likely to succeed. For example, it is not impossible to raise support for socialist reforms within a capitalist economy, but the fact that the economy is capitalistic, and the extent to which people support the fact that it is such, will make socialist reforms more difficult to implement, and therefore less likely to succeed. Likewise, it is not impossible to introduce reforms that go against the culture, or the religion, of the citizens in some society, but we can expect much less compliance, and much more resistance, when that is the case. In some instances, if we want the reforms badly enough, we will have to be prepared to really manipulate people's incentives in order to secure success. To a certain extent, accepting soft constraints does mean accepting that the status quo places a limit upon what we can realistically accomplish. But conceiving of the limitations as soft constraints rather than hard ones emphasizes the fact that we can work around them. For example, we might think about how we can introduce changes that will gradually erode the soft constraints, so that at a future time, they will not be constraints at all. It is one thing to use morally reprehensible forms of coercion to abruptly introduce reforms that clash with deeply-held religious beliefs, it is another to increase education in e.g. the natural sciences in order to gradually draw people out of their religious stupor.

5.4.2 Psychology and motivation

More difficult to characterize are the constraints of individual psychology and motivation. Should we take *everything* about the status quo to be a constraint, or are there some facts we should rightly ignore? Motivation seems like the strongest candidate for ignoring, especially given discussion about 'ought implies can'. The fact that a person *won't* do what he ought is no reason to say he is not required to do it. The fact that the citizens in some society won't support reforms that have been proposed does not entail that they are not *required* to support those reforms. If there is an obligation or a requirement, it must be that there is at least binary feasibility. The question that remains is whether we can get from that to the idea that people's being unmotivated or

psychologically such that they are unlikely to act in certain ways counts as a soft constraint upon a given action being likely to produce a given outcome.

David Estlund (2010) talks about motivational and psychological constraints, introducing the case of Bill, a man who constantly fails to deal with his rubbish in the required way, instead just dumping it in the street. Bill understands that he ought to deal with his rubbish in the proper way, but just cannot bring himself to do so. This is a failure of action, but it comes from a failure of will. Estlund's question is whether we should think that failures of willing are sufficient to warrant a judgement of inability, whether they dissolve obligation or absolve responsibility (see also Jackson & Pargetter, 1986, discussed in more detail in Chapter 7).

Consider a few alternative examples of failure to will. You might trip over and smack your head on the pavement, but it seems that you can't have really willed for that to happen. You might walk an incredibly detailed route home, thinking about all manner of things, making all sorts of small idiosyncratic bodily movements, without having been able to will that trip in all its intricate detail. You might visit your friend who works at a charity and accidentally leave a wallet full of money behind, without having willed making such a large donation. These are all cases in which a person fails to will an action which they can nonetheless do. But it seems that any plausible case of physical ability and psychological inability, or psychological ability and physical inability, relies on equivocating between the two. You are able to leave a large donation at the charity where your friend works; we think that's true because we concentrate on the physical case and bracket the psychological case (you never would have willed to do that, but you did it). Or we might instead say you're *not* able to make a large donation to the charity; we think that's true because we concentrate on the psychological case and bracket the physical case (you left the money accidentally; it wasn't an intentional action). But whatever features of a situation are relevant to determining ability should be part and parcel of our judgement about the situation. If it's true that you can make a charitable donation, that's partly because you can will it, and if you can't will it, that is a reason to say that you can't do it.

Of course, Estlund's question still remains, which is what we should say about Bill, who ought to put his rubbish out correctly but fails to will it, or the charity case, in which you ought to make a charitable donation but you just can't bring yourself to do it. Notice that there are obvious psychological limitations which do seem to make some action unlikely to produce a given outcome, such as addiction, or compulsion. If my

question is how likely Bill promising his neighbours that he will reform his rubbish-related behaviour is to result in the outcome that he actually reforms his behaviour, and I know that Bill has a severe compulsion toward dumping his rubbish in the streets, then the answer should be 'not very likely'.[31] But that is not the kind of case at stake here; at least with Bill, 'there is no specific phobia, compulsion or illness involved'. Thus despite Estlund's claim that Bill 'wishes he had more willpower' (Estlund, 2010, p. 10), the description in terms of Bill's failure to bring himself to do what he ought seems to be no more than fancy talk for the fact that he *doesn't really want to do what he ought*. Perhaps there is some second-order sense in which he does, i.e. he wants to want to. But clearly, Bill has a set of options available to him, and from the fact that he chooses to dump his rubbish in the street we can infer that he preferred, all things considered, dumping his trash and taking whatever perceived benefits that conferred upon him, to dealing with it correctly and accepting whatever perceived burdens that conferred upon him. In the binary sense, used to rule actions out of Bill's option set, if his preference were severely pathological we might dissolve his prima facie obligation. But failures of will are not sufficient to binary infeasibility. Estlund agrees with this conclusion, arguing that "ought' does not imply 'can will" (Estlund, 2010, p. 4). But are they sufficient to *less* feasibility? If I am thinking about whether a small group of committed neighbours can introduce some environmentally-beneficial changes to rubbish management, should I take as a constraint upon what they can achieve the fact that Bill is likely to keep behaving as he does? It seems to me that I should. From my perspective, Bill is a static part of the environment which bears on whether the actions the neighbours choose can realize the outcomes that they desire. Bill is unlikely to deal with his rubbish properly; if that affects the likelihood of his neighbours' actions producing their intended outcome, then it affects whether that outcome is feasible. Notice that this is not the same as saying that Bill is not required to deal with his rubbish correctly. That is a separate matter. If he both ought, and it is an option of his that he do, then he is blameworthy for failing to. But which moral obligations bear which which strengths on which individuals is a separate question from what it is feasible for an agent to do *given* other people as part of the constraints of his situation.

We should take motivational and psychological constraints seriously when we think about feasibility in this second, graded role, but only when thinking about other

31 Probably it is best to understand pathology as a continuum of behaviour. The greatest amount of pathology will rule out an outcome following from an agent's (trying to) act, the least amount may have no affect at all.

agents who partially constitute the soft constraints bearing on my action producing the desired outcome. It does not matter whether I am likely to choose an action in my option set (more on this in Section 5.4.4) but it matters whether my action is likely to succeed in producing the outcome I intend, given motivational and psychological facts about other people. It is important that we take seriously *what other people are like* when we think about what we can do, and to what degree we can do it.

One worry is that there might be feedback loops from our taking seriously some agent's being unlikely to act to his being let off the hook, because the outcomes I can achieve are part of what create his obligations. Imagine that I am one of Bill's neighbours, and I form a collective with our other neighbours, with a particular scheme in mind for recycling and waste management. We in the group have a choice about whether to include Bill as someone whom our demands bear upon, or to exclude him. If we think there's little we'll be able to do to convince him to deal with his rubbish appropriately, then we might just leave him out of our scheme. But now it looks like Bill, through his own deliberately chosen actions (he *could* put his rubbish out correctly, but he *doesn't*) is 'off the hook' with respect to our rubbish and recycling scheme. He doesn't have any obligations because we have deliberately left him out of our plans. But is that right? Should it really be that people benefit by being stubborn and dogmatic about their own habits and their own character, even when change is urgently morally required? I have to reiterate that this is a separate question to questions about what is comparatively more feasible. Our action when it includes Bill has little chance of succeeding, because of Bill's habits, and that is what justifies us in judging the action to be less feasible than another alternative that doesn't include Bill. What Bill is obliged to do, whether he is off the hook too easily, and whatever his secondary obligations might be in light of his failure to realize his primary obligations, are all moral questions, and as must be clear by now, I have largely set those questions aside here.

In fact, I think the question about how to characterize psychological and motivational constraints is something of a red herring when it comes to feasibility in the comparative role. It's true that we don't want to capitulate to a kind of cynical realism about what is feasible, and it's true that we don't want to let people off the hook when they fail to do what they could easily have done. The crucial difference, however, between the moral philosophical project and my project (at least in this chapter, for this sense of feasibility), is that we can't infer anything from requirement to action. Moral philosophers want to know when facts about what a person can do are sufficient to make

it the case that the person is not required to fulfill a prima facie obligation. But their concern ends there. Mine does not. I might be interested in the same question, but I want to know more: not only what people are required to do, but what they are *likely* to do, what we can reasonably expect them to do, what they can do if they try, and so on. If people lack the motivation to φ, we might think a proposal requiring them to φ will be less successful than a proposal that doesn't require that. It's obvious that psychological and motivational constraints are *soft* rather than hard, because they can be manipulated and they can be overcome, with varying degrees of difficulty. What's still an open question is whether we should think of them as constraints *at all*.

The answer to this question of whether we should include motivational constraints in the package of soft constraints at all, foreshadowed earlier in this subsection, is that it depends on whether we're talking about the agent who has some action available to her that could realize the desired outcome, or about other agents who constitute part of the context in which she acts. (This will be important as the thesis goes on in Chapter 7 to problems specific to collective action). I think psychological and motivational constraints are relevant to the context our agent acts in, but not to whether our agent chooses the action in question. When we think about what we can get others to do, or what we can expect from others, we should take them as they are, and if they are excessively unmotivated, it will be prudent to factor that into our assessments. But I should not, when I am thinking about what I can do, think about whether I am likely to do it (or try to do it), and the same goes when assuming the perspective of another person whose project we are assessing for feasibility. I follow Geoffrey Brennan and Nicholas Southwood on that point, which they introduce through the following case:

> Suppose that one is so lazy that one is highly unlikely to go to one's daughter's hockey game on Saturday morning. It's not that there is anything preventing one from doing so. Nor would it be especially difficult or costly. It's just that one is so lazy that one will almost certainly stay in bed and read the paper instead. It would be a mistake, we take it, to claim that one's going to watch one's daughter's hockey game is "infeasible". It's perfectly feasible. It's just that the chances are that it won't happen. Feasibility isn't the same as sufficient probability. (Brennan & Southwood, 2007, p. 9).

Brennan and Southwood want to show that feasibility and sufficient probability come apart, because there are many things that it would be feasible for a person to do that she is nonetheless unlikely to do. The parent in their example is unlikely to go to his daughter's hockey game, but there's nothing stopping him doing so. None of the soft constraints mentioned so far make the parent unlikely to succeed in going to the game,

not economic, institutional, or cultural constraints. And he has no pathology (or none that the authors mention) that makes him psychologically unlikely to succeed in going if he chooses to (tries to); the only thing is that he is *lazy*. He prefers to stay in bed and read the paper. The question is how we should treat this kind of case. And as I have said, I think the answer depends on whether the lazy parent in Brennan and Southwood's example is the agent or part of the context in our assessment. That's because I think that we are justified in assuming the motivation of those whose projects we are concerned with. Let me try to motivate this claim.

Suppose my question is 'is it feasible for me to get Olivia's father to come and watch her hockey game?' Then what I am asking is whether the relevant soft constraints make it more or less likely that I will succeed in getting Olivia's father to come and watch her hockey game. I can assume that I will try to get him to do so:

> In the first person case, in deliberation, a rational actor does not consider a possible action irrelevant because of her own unwillingness to perform it, but instead considers the full range of things she could do. ... In deliberating about whether or not to do *X* at *t*, or indeed about whether to perform an extended sequence of actions of which doing *X* at *t* is a part, the rational actor *ignores* any predictions she might have about whether she will do *X* at *t*. This seems to be a presupposition of rational deliberation (Woodard, 2003, pp. 224-225).

The big question, for me, is not whether I will choose that action, but what my chances of success are. If Olivia's father is part of the context in which I act, it makes sense for me to factor in his laziness. Whether I will succeed depends on exactly how lazy he is. But suppose my question is instead, 'is it feasible for Olivia's father to go and watch her hockey game?' Then *I* am no longer the agent of interest; Olivia's father is. And in that case, I assume that *he* will choose to go, and I think about whether he is likely to succeed. Given that the only thing standing between success and failure is his lack of choice, I will probably judge that it *is* feasible.

This demonstrates the fact that we get different answers about what is feasible depending on how we partition agents, whether it is them whose actions we are assuming can produce the desired outcome, or whether they are part of the context in which another agent acts. The pattern is the same when we increase the scale. Suppose we are wondering, for any particular dog-owner in Berlin (and there are many), whether it is feasible that she clean up after her dog when it excretes in the streets of the city. We probably judge that she can, because there are no soft constraints making it unlikely that she'd succeed if she chose to do so. But suppose we are instead wondering from the perspective of the Berlin city council whether it is feasible that they introduce new

bylaws requiring dog-owners to clean up after their dogs, or whether it is feasible that through introducing a system of fines they get the dog-owners to clean up after their dogs. Then the question is whether their action will bring about the outcome, in other words, whether introducing new bylaws and fines will cause dog-owners to clean up after their dogs.

The fact that we get different answers about what is feasible depending on the context is not contradictory. Usually, we will ask feasibility questions from the perspective of those whose project is at issue. If we're wondering about introducing reforms with respect to dog excrement in Berlin, our agents will be those who have the power to introduce such reforms. Certainly it is true that if we ask what is feasible for Queen Elizabeth we might get an answer that something is feasible even though *it will never happen*. But there are two things to say about this. The first is that it can be useful to know what people *could do if they chose to*, that for example any dog-owner in Berlin could comply with the proposed reforms if she chose to. The second is the old adage, 'ask a stupid question, get a stupid answer'. If we ask what is feasible for a person with whom we cannot communicate, over whom we have no influence and absolutely no reason to expect to *do* the actions we're interested in them doing, it should be little wonder that, even though the action in question may be a highly feasible means to an outcome, the agent in question doesn't end up doing it. This doesn't mean the outcome *isn't feasible*, it just means that the point of thinking about what is feasible is often, in the end, to combine various other considerations (more on which in Chapter 6) in order to make decisions about action. The merely theoretical finding that something is feasible doesn't get us the whole way, on its own, to *what will, or should, be done*.

5.4.3 Positive Morality

While I think we should refuse to admit moral constraints into the package of soft constraints that bear on the purely theoretical question of what is feasible, it seems clear that we should include the constraints of positive morality. The fact that some action is *immoral* is not sufficient for its being infeasible. There are lots of actions in the history of the world that were immoral and nonetheless ended up being done. Just think of King Leopold of Belgium in the 18th-19th Century, who demanded that his soldiers either kill or chop the hands off indigenous people in the Congo Free State, in one of the worst atrocities of the century (Hochschild, 1999). I should hardly need to give

further examples; they are everywhere. And if they are actual, then *a fortiori* they are feasible. If an action is accessible to a person, then what matters is whether the soft constraints of the context she is in make it more or less likely that the action will succeed in producing the desired outcome. That is a causal question, not a moral one. But while moral truths alone do not make an action more or less likely to succeed, people's *beliefs* about moral truth might do so. That is to say, there are feedback effects between positive morality, the moral beliefs people in a particular society happen to have, and what it is possible to achieve in that society. Let me give a couple of examples of these kinds of feedback effects.

First of all, if people are convinced that a certain outcome is extremely desirable, let's say because it accords with their deeply-held religious or cultural views, then the outcome may for that reason end up *being* more feasible, simply because people will pursue it much more vigorously than other alternatives. Second of all, the fact that people have certain moral commitments might cause them to be more serious in their thinking about what is feasible and what is not. In thinking more clearly about the options available to them, they may reach more accurate assessments of what is feasible, and thus be able to pursue options that with less serious enquiry would have seemed out of reach. Thirdly comes the problem of adaptive preference formation. In this case, agents modify their desires to fit their ideas about what is feasible. This can be pathological, because it means that if a person has inaccurate ideas about what is feasible, she might thereby fail to do as much as she otherwise would, in the pursuit of genuinely desirable (from her prior perspective) goals. Finally, especially from a policy-maker's point of view, a reform might seem more desirable simply because it is more feasible. Thus there is feedback in both directions, from positive morality to feasibility, and back again. And unlike moral constraints, we do need to factor these effects, and positive morality in general (although much of this is subsumed by culture and religion), into the package of soft constraints, because people's beliefs in this respect will impact upon the degree to which a given reform is likely to succeed in a particular context.

5.4.4 Effort

Another soft constraint upon an action's producing an outcome is the amount of *effort* required by those in the context in which the action takes place. There's probably no real line at which no greater effort can be made in achieving some goal, but the more

difficult and demanding some action is, the less likely a person is to do it, unless she is unusually committed to the cause requiring it. That is a platitude about human action. This is a constraint that moral philosophers often try to capture under the heading of 'over-demandingness'. For example, Peter Singer proposes an account of our charitable obligations which basically requires that we give until we have almost nothing left for ourselves. Many people object to his account on the grounds that it *requires too much of people*, that is it unrealistic in the demands it makes. It is not that anything in particular rules out people giving that much, it is just that people would have to try really, really hard to conform to his requirements (or perhaps just to convince themselves that they *should* conform to his requirements) (Singer, 1972; 2009).

Let me introduce a case to make an important distinction. Clas is solely concerned to win the affections of Astghik, and because he knows that her sole passion in life is opera, he decides that putting on an opera would give him the best chance of realizing his goal. He could just buy her some flowers and some chocolates, but he thinks this would be far less likely to result in her returning his affections. A grand gesture is required. For simplicity imagine that Clas has only these two options available to him: put on a magnificent opera, or buy some flowers and chocolates. He assigns a far greater likelihood to the former than the latter in realizing his aim, namely winning Astghik's heart. But now imagine that it would be *really, really difficult* for him to put on an opera. He knows little about opera and he isn't particularly good at organizing things. He also has limited resources. He'd have to use all of his money, all of his time for the next few months, and his most ingenious efforts at persuading others to help him, to pull the opera off. On the other hand, it would be *really, really easy* for him to buy some flowers and some chocolates. He has more than enough money to manage this.[32]

Here we're not talking about other agents in the context in which our agent's action takes place, but about whether the extent of his own effort is a constraint on which outcomes are feasible for our agent. The intuition that some might have here is that it's less feasible for Clas to put on the opera than to buy the flowers and chocolates. But I think we have to resist that intuition. It rests on a conceptual confusion between *easy* and *hard*, compared with *feasible* and *infeasible*. The natural thought is that things that are easier are more feasible, and things that are hard are less feasible. But these come apart. Some things that are really hard are feasible. Imagine that the future of humankind hangs in the balance depending on whether Clas can put on the opera

32 I am grateful to Clas Weber for discussion on this point.

(perhaps it does; perhaps the child of Clas and Astghik will discover a cure for a super-virus that threatens to wipe out the whole population). He *can* put the opera on, it would just stretch his resources. If there's enough riding on it, then it looks like putting on the opera is exactly what he should do. So it's really hard, but it's still feasible.

On the other hand, something might be really easy but infeasible. Imagine that all I have to do to double my money is go to the casino and put all my money on black. Going to the casino is really easy, and so is putting all my money on black. But the chance of my doubling my money is slightly less than half (the odds are in the house's favour). So we might judge that the action of putting all my money on black is an infeasible way to produce the outcome of doubling my money. There are plenty such examples. The point is that we shouldn't confuse easy and hard with feasible and infeasible. We can admit that it would be really hard for Clas to put on an opera for Astghik, but we can still insist (a) that it is feasible for him to do so, in the binary sense that it is one of his options, and (b) that what is interesting *for comparative feasibility* is only the extent to which that action is likely to produce the outcome (the opera is likely to result in his winning Astghik's affections). It is of course relevant to what Clas should do that it would be really hard for him to put on the opera, but as I have said already, choice-worthiness and feasibility come apart, which is why feasibility supplements decision theory without being able to supplant it.

5.4.5 Soft constraints, summary

To recap, I began this chapter with the claim that a proposal is more feasible the less it clashes with the relevant soft constraints. Some of these are part of the environment in which an agent acts, such as the economic and institutional constraints; others come from the agents who partially constitute the context in which an agent acts, such as the constraints of culture and religion, psychology (stronger toward the pathological end of the spectrum) and positive morality. In that case we take people just as they are. When we ask about any agent in particular we are licensed in assuming effort, when we ask about agents partially constituting the context in which she acts we are not. Let me try to put this in terms of comparative feasibility tests, to complement the binary feasibility tests presented in Chapter 4:

(General) Graded Feasibility Test: The degree to which the recommendations of a theory are feasible is established by the conditional probability of the recommendations being realized *given* an action (or set of actions), where the action is in the option set of some agent, and where the action is the most likely of any action available to any agent to bring the outcome about.

(Agent-relative) Graded Feasibility Test: The degree to which the recommendations of a theory are feasible for an agent is established by the conditional probability of the recommendations being realized *given* the action (or set of actions) most likely to bring the outcome about available to the agent.

The logic of the graded feasibility test incorporates the 'given that' operator. This is not a conditional ('if, then') but a conditional probability. What is important is the probability of the outcome given the action, i.e. the probability that the outcome will be brought about assuming that the action is performed. In the English language it is hard to distinguish conditional probability from counterfactual probability, e.g. the probability that the outcome would obtain given that the action were performed. Decision theorists have found all sorts of problems in doing things with conditionals, and they have resolved them by turning to counterfactuals with imaging (see e.g. Hájek, 2002). If my use of conditional probabilities turns out to be beset by similar (or simply irremediable) problems, then I can also exchange the conditional probability in the graded feasibility tests for counterfactuals with imaging. But for the time being, I cannot see any such problems, so I shall leave them as they are.

In this formulation we focus on how feasible some outcome is by asking, for some agent, how likely her action, feasible in the binary sense, is to produce the desired outcome. We simply assume that she can choose to perform that action if she wants to; it is an action in her option set. *Whether* she will choose it is a whole other story, and irrelevant for the comparative test of feasibility. The recommendations of a theory are feasible to degree *X*, and *X* is established by taking the action, feasible in the binary sense for some agent, that has the *highest* probability of producing the outcome, and looking at *how likely it is to produce that outcome*. The likelihood is determined by looking at the effect the soft constraints have upon the action's producing the outcome. The more likely success is, given the relevant action, the more feasible a proposal is; the less likely,

given the relevant action, the less feasible.[33]

5.5 Running the Comparative Feasibility Tests

In Section 4.3, I introduced Carrot Mob, the Berlin group who want local businesses to switch to more energy efficient appliances. I made the fairly innocuous assumption there that it is obligatory that businesses in general make this kind of switch. Consider (P2) again, 'it is feasible for Carrot Mob to get *Krawall & Remidemi* to change their appliances to more energy efficient ones'. This proposition is not about obligation, but about feasibility. It is a proposal about *one way in which the obligation might be realized*. Mostly we think that obligations belong to those they bear upon, which is to say, the responsibility for a person's fulfilling their obligations rests only with that person. But in our non-ideal world, not all people are motivated to fulfill their obligations. Then if the choice is between a world in which more obligations are fulfilled because more people take responsibility for others' fulfilling them, and a world where less obligations are fulfilled because people's choices with respect to their own obligations are theirs alone to make, we might reasonably think the former world is the better one (although the tradeoff between moral superiority and greater autonomy in a world is certainly not insignificant). How do we figure out whether that proposition about feasibility is true? Is it really feasible that Carrot Mob succeed in getting *Krawall & Remidemi* to switch its appliances?

There are many things to consider in answering this question, but they have all been introduced already. The first step is to reintroduce the binary sense of feasibility developed in Chapter 3. Is *Krawall & Remidemi* switching its appliances ruled out by any hard constraints? Answering that question requires answering two further questions. Is (P2) made false by any timeless hard constraints, such as the laws of logic or nature, or the extent of our biological knowledge? Clearly it is not. Does the action or outcome in (P2) violate any time-indexed hard constraints, such as the limits of technology, or a fixed budget, or available resources? It seems that the answer here is also 'no'. The cafe may not be able to afford the change at the time they are approached, but part of the

33 A further question we might ask with respect to soft constraints is whether any have greater strength than any other. For example, is it worse for the recommendations of a theory to clash with cultural constraints than it is for them to clash with economic constraints? Or, is it not such a big deal if a proposal requires a lot of effort, but a big deal if it requires people to do what they are not interested in doing? At this stage I simply leave it at saying that the constraints are all important and will all impact on the likely success of a proposed action, but it would be interesting to think more about whether constraints should instead be ordered in terms of relative priority.

Carrot Mob project is to cover the costs of their making the change. Thus the proposition is true according to the binary test of feasibility. The second step requires applying the comparative sense of feasibility introduced in this chapter. To what extent is Carrot Mob likely to succeed in getting the cafe to switch their appliances, given that it chooses to (try)? The likelihood that it will succeed in producing that outcome given its action determines how feasible the outcome is. What exactly do we need to know to answer that question?

First of all, we need to know about the actions Carrot Mob proposes to take to bring about its desired outcome. Given the description of how the group works, we can assume that to get the cafe to switch its appliances, Carrot Mob would have to put its efforts into a campaign to get people to go to the cafe on an agreed day. The group might put advertisements on the local radio station, distribute fliers giving people information about the date, venue, and project, talk to people in the streets about what they can do to help, and so on. Assuming that this is roughly the best available action, we have to ask how likely it is that the action will produce the desired outcome. Without giving an exact probability, it seems we can be fairly confident that these actions would be efficacious in getting people to visit the cafe (similar actions on the group's part have been successful in the past). The more people that visit the cafe on the agreed day, the higher the cafe's profits above normal, and thus the more likely the project is to succeed, i.e., Carrot Mob get *Krawall & Remidemi* to switch its appliances over.

I cheated a bit in the last paragraph by inferring likelihood of success from known past successes. Let's pretend that we don't have access to any such information. Then we might want to ask about the truth of the further proposition: 'it is feasible that a sufficient number of people respond to the Carrot Mob campaign and visit *Krawall & Remidemi* on the right day.' How do we assess *this* proposition? It is an important question, because it taps into the problem that at bottom, the success of most collective projects will depend on the motivation of individual persons. If sufficiently many people respond to the Carrot Mob campaign, which is to say, are motivated to go to the particular cafe on the particular day to serve that particular political goal, then the project will succeed. If they are not, then it will likely fail. But we have defined feasibility in the comparative sense as conditional upon trying. Surely if people try, they will succeed? There seems to be no reason to think they would not; nothing rules it out, and nothing even seems to bear on making it less likely. Even if we think 'difficulty' (i.e. approaching the limits of available effort) is a soft constraint, there's nothing especially difficult about going to a

cafe (especially in Berlin, where a large proportion of the population are artists of some kind, who largely determine their own schedules). But remember that we must be careful about assigning people to the context in the right way. We are considering a project of *Carrot Mob's*, so they are the relevant agent. All those who they would seek to convince to visit the cafe are part of the context. How likely those members are to be convinced will determine whether the project will succeed or fail.

To restate this point, we must always talk about feasibility for a given agent. Consider again our three different proposals. (P1) It is feasible for *Krawall & Remidemi* to change their appliances to more energy efficient ones. (P2) It is feasible for Carrot Mob to get *Krawall & Remidemi* to change their appliances to more energy efficient ones. And (P3) it is feasible for Clas to respond to Carrot Mob's campaign by going to *Krawall & Remidemi* on the required day. On my account, (P1), (P2) and (P3) are true so long as the relevant agent's (*Krawall & Remidemi*, Carrot Mob, and Clas, respectively) action (changing their appliances, getting the cafe to change its appliances, and going to the cafe, respectively) would be likely to produce the desired outcome (having more energy efficient appliances, causing *Krawall & Remidemi* to have more energy efficient appliances, and being a part of causing *Krawall & Remidemi* to have more energy efficient appliances, respectively).

5.6 Feasibility and likelihood, two distinct issues

Feasibility has a special subject matter. It is not merely about what is *possible*, but neither is it merely about what is *likely*. In all of its roles, it tries to occupy a space in between those two extremes. Instantiating political theories and realizing political recommendations must be possible (in the philosophers' sense), but it must be more than that, too. It must not be so much more, however, that all theories and recommendations do is recommend what is easiest, or what would most likely eventuate without the theories and recommendations. Non-ideal political theories play a certain role; they are *aspirational* without failing to be *action-guiding*. But a critic of the concept of feasibility might ask why we should be interested in feasibility at all. Especially in its forward-looking role, feasibility assessments are instrumental in deciding which outcomes to pursue. Pursuing outcomes costs time and money; that's why we're interested in only pursuing the ones that have some decent chance of being realized, should we actually aim at them. But then the question is: why care about what is feasible? If we just

concentrate on what is likely, we won't make any mistakes in deciding on policy and institutional reform.

The answer to this question is that likelihood is relevant for feasibility, but only at a certain point. It is *irrelevant* to feasibility when we are restricting our attention to feasibility *for* some agent what that agent is likely to do. Leone is prima facie obliged to finish her homework so that she can keep her promise to meet her friends. But she is not obliged if we can show that she has no actions in her option set that could produce the outcome of her finishing her homework. And she may be less obliged (depending on what is at stake) when an action in her option set is not *very* likely to produce the outcome of her finishing her homework. But which actions are in her option set, and how likely they are to bring about the outcome that she finishes her homework, is not in the least bit affected by how lazy or useless she is. The fact that she probably won't *choose* a certain action, or *try* to perform it, has no effect on the fact that the action is in her option set. For the binary feasibility test that is all that matters: whether the action is ruled in, or whether it is ruled out. For the comparative feasibility test, all that matters is how likely the action is to bring about the outcome. *That* is the sense in which mere likelihood is irrelevant to what is feasible *for* Leone.

The sense in which likelihood *is* relevant to what is feasible for Leone is that what other people are likely to do can have an effect on whether her action is likely to bring about a certain outcome. Assume that she chooses to stock up on coffee and chocolate, and spend the night in working on her assignment. If she succeeds in doing that, she will finish the assignment in plenty of time to be able to meet her friends in the morning. And imagine that, uninterrupted, the fact that she chooses that action would result in her performing that action, and the fact that she performs that action would result in the outcome that she finishes her assignment. But now imagine that her flatmates come home drunk, celebrating the recent engagement of someone in the philosophy department, and forcibly disentangle her from her work to join them. The fact that her flatmates provide a stubborn obstacle to Leone continuing working is relevant to whether her action succeeds in bringing about the outcome. In the absence of interference it would have, but her flatmates are one of the environmental soft constraints upon her action that make it less likely to produce the outcome in question.

To summarize, thinking *only* about what is likely is counterproductive, because political theory rests on the assumption that change is possible. But we should not assume that everyone is perfectly virtuously motivated, because that will tell us that a lot

is feasible which may actually not be. The trick is to talk about feasibility *for* particular agents. If we just ask whether an outcome is feasible, it is a matter of whether *any* agent has an action that would realize the outcome. But if we ask whether an outcome is feasible for a particular agent, then we can assume her motivation, and look at what could, and would most likely, follow from that. Any action in her option set is up for grabs. Then it is *conditional probability* not *probability* alone that is important. Graded feasibility is about what is likely *given some action in the agent's option set*.

5.7 Conclusion

Accepting the account of graded feasibility given in this chapter will force us to say that something is feasible to a high degree even though *it might never happen*. I have said that I don't think this is the wrong result, because we conditionalize upon only the motivation (or action choice) of those whose potential outcomes we are assessing. I have argued that graded feasibility, which allows ranking and comparisons between various recommendations, is a matter of how likely some action in an agent's option set is to bring about a desired outcome. Once we can rank, we can introduce the further elements necessary for choosing which outcomes to pursue.

In the next chapter, I want to ask whether feasibility is epistemically tractable. Obviously, what is true comes apart from what we (can) know. Perhaps there is a fact of the matter about whether a given proposal for reform is feasible in the binary sense, and the degree to which it is feasible in the comparative sense. But in terms of making decisions, we can only go on the best evidence we have available. In the later part of the next chapter, I ask just how good we are in assessing the counterfactuals necessary to feasibility claims. In the earlier part, I ask how we are to combine feasibility with desirability considerations, and epistemic limitations, in order to reach all things considered practical judgements about what, politically, we ought to do.

Chapter 6

Feasibility in Practical Reasoning

6.1 Introduction

In the last two chapters I introduced two tests of feasibility, the binary feasibility test, in which the recommendations of a theory are either feasible or not, made infeasible by violating either timeless or time-indexed hard constraints, and the comparative feasibility test, in which the recommendations of a theory are feasible to a certain degree, and can be ranked in a set of more and less feasible recommendations. By now we should have some sort of handle on what the concept of feasibility amounts to. But that is not the end of the story. Having a theoretical grip on the concept does not tell us how to utilize it in practice. In this chapter I will argue that two further elements are required before we can use feasibility to reach all things considered judgements about what to do. The first is an epistemic element, introduced in Section 6.2, the second is a normative element, introduced in Section 6.3. Discussion of the epistemic element will be framed in terms of an objection to the formulation of feasibility defended thus far. I have said that an outcome is feasible in the binary sense if there's an action available to an agent that could bring it about. But I've assumed a fact of the matter about which actions are in the option sets of agents. Shouldn't we instead impose an epistemic constraint and say that an outcome is feasible if the agent *knows* there's an action in her option set that could bring it about? I will argue that we should resist this epistemic constraint, but will concede that there's an epistemic constraint on those assessing feasibility, in that assessments are relative to best available evidence.

Discussion of the normative element is motivated by the idea that decisions about which political reforms to implement in the real world must come from tradeoffs between what is feasible and what is desirable. In Section 6.4 I will introduce basic decision theory as a way to formalize these three important elements of practical political decision-making. As mentioned in Chapter 1, two roles feasibility can play are as supplements to decision theory. One allows us to say more about how a judgement of non-zero probability for an action's being done by an agent, and an outcome's being produced by an action, and a judgement as to the probability of an outcome given an action, are reached; another is to act as a decisional heuristic when the full decision theoretic calculus is not available. I suggest that because what is feasible comes apart

from both what is likely, and what is actual, assessing it requires a special skill: the ability to assess counterfactuals. We judge that one set of recommendations is more feasible than another because we judge that the action(s) most likely to bring it about are *more* likely to bring it about than the action(s) most likely to bring about the other. Reaching such a judgement requires thinking about what would happen *if* the agent chose to perform the action, available to her, most likely to bring the recommendations about. Just how good are we at assessing counterfactuals, though? If we are terrible, the whole project is in trouble. That question will be the focus of the last section in this chapter, Section 6.5.

6.2 The epistemic element of feasibility

In Chapter 5 I mentioned that some people might be dissatisfied with the way the comparative feasibility test conditionalizes upon an action. Whenever we assess feasibility *for* a particular agent, we assume their doing an action in their option set. But that means that sometimes some action will come out as feasible, even though it is highly unlikely to ever be undertaken (see also discussion in Section 5.4.4). In this section I want to outline the possibility of limiting the account of feasibility to what agents *take* their options to be, and explain why that move should ultimately be resisted.

With respect to Brennan & Southwood's (2007) case of the father who is too lazy to go to his daughter's hockey game, I said in Section 5.4.2 that what we should say depends on whether the father is the agent whose actions we are considering, or part of the soft constraints that characterize the context in which another agent acts. If the father is the agent we're concerned with, then we'll judge that it is feasible that he go to his daughter's game, because none of the relevant soft constraints make him unlikely to succeed in going if he chooses to. Because we want to know which outcomes are feasible *for the father*, we conditionalize on the actions that are available to him. But if the father is part of the soft constraints, for example in the case that I am thinking about whether *I* can persuade him to go to his daughter's game, then I may well judge my action to be infeasible, because despite my best efforts, his laziness might be immune to my persuasion.

One reason to think Brennan and Southwood say the right thing about the hockey-father case, namely that it is feasible that he go to his daughter's game, is that we generally assume the father in their case is *aware* that going to his daughter's hockey

game is one of his options. That is to say, the option is salient to him.[34] We might assume that he feels the normative force of that option, knowing for example that it is important for parents to support their children in their various endeavours. Because we make these assumptions about what the father takes his own set of candidate actions to be, we judge that it is reasonable to conditionalize upon certain of the actions available to him. We know that people can act in certain ways *if they choose to*, even if we also know they won't choose to. That is to say, choosing, or 'effort', or 'trying', are pre-screened for binary feasibility. But perhaps these assumptions about an agent's epistemic or doxastic states (knows he could do the action, or believes he could do the action, if he chose to) are what close the gap, at least somewhat, between feasibility and likelihood. An outcome is feasible because the agent knows there's an action available to him that could bring it about. If he didn't know an action was available, we could certainly not expect him to choose it, and then could not expect the outcome to be brought about. Surely this is a preferable sense of 'feasible'? We can test this conjecture by thinking about another case, in which an option is *not* salient to the relevant agent.

Imagine instead that the father in Brennan & Southwood's case has been long since estranged from his family, and has no idea that his daughter is playing a game of hockey today in the city where he lives. From his own perspective, imagine that his options for the day are 'stay in bed and read the newspaper', 'go out and walk the dog', and 'pick up some groceries and cook something interesting for dinner'. He doesn't *know* that there's a hockey game, and therefore he doesn't feel any kind of normative pressure to go to it. Similarly, imagine that Angela Merkel is thinking about the feasibility of various policy arrangements between Germany and the United States. What *President Obama* can do if he tries is different to what *Chancellor Merkel can get President Obama to do if she tries*. In the former, Merkel thinks about what is feasible for Obama, so she may assume his trying. In the latter, Merkel thinks about what is feasible for Germany, given the United States as one of the soft constraints upon the success of its actions. She might think about what he is likely to do, or what he can be brought to do given *her* best efforts.

But why should we be interested in Merkel's thoughts about what Obama can do if he tries? After all, those results are completely uninformative if she has no reason to

34 It is not clear to me exactly how this should be cashed out, although ultimately it won't matter because I will reject the need for such a condition. But if it were needed, a choice would have to be made between saying that candidate options are those (a) a person has thought about doing (the problem with this is that a person might entertain some action but have no serious intention to do it), and (b) a person takes to be a 'serious' option of theirs in some sense (then we'd have to define a threshold past which an option is taken seriously enough).

believe he will actually choose to do any of the actions she imagines him doing. Obama might not choose any of the actions Merkel assigns to his option set. What outcomes he can bring about (and with what chance), given he does an action in his option set, is a different question to what he can be *expected* to do. Do we really want to say that it is feasible for the estranged father to go to his daughter's hockey game? And do we really want to say that it is feasible for Obama to do what Merkel judges it feasible for him to do?

In discussions over 'ought implies can', you do not often hear people say that what a person can be obliged to do is limited to what she *believes she can do*. The limits to obligation are discussed in ontological, not epistemic, terms. But clearly there will be cases in which a lack of knowledge is sufficient for an excuse in failing to act. Imagine I believe that I cannot swim, and this belief is justified by the fact that I never learned. And imagine that I do not attempt to rescue a drowning child, because of my belief. If, unbeknownst to me, it turns out later that I had a natural ability (if there is such a thing) to swim, surely few people would want to say that I am guilty for my past failure to act. I was ignorant of certain facts about myself, but I was non-culpably ignorant. If neither the estranged father nor Obama take certain actions to be potential actions of theirs, then we cannot expect them to do them. And if we cannot expect them to do them, why should we say that they are feasible actions?

The answer is that there is a difference between *what a person can be expected to do* and what a person *can* do. If the hockey father doesn't know there is a hockey game, we shouldn't put much credence in the proposition that he'll go to it. But there's still a separate question about whether it would be feasible for him to go to it. The gap between feasibility and likelihood remains vast, and that seems like an important feature of feasibility that we should aim to preserve. It might not be very interesting to know that the father could go to the game if he chose to, seeing as we can in all likelihood expect him not to choose to (because he doesn't know about it). And it might not be very interesting to know what (Merkel thinks that) Obama can do if he chooses to, seeing as there's no reason at all to expect him to choose it. But if we want to know about what the estranged father or Obama are likely to do, it is not as though the comparative test of feasibility has nothing to say. What it recommends is that we take them as part of the soft constraints of a context. When we do that, we take people just as they are. What the estranged hockey father is likely to do is a constraint on what I can persuade him to do. What Obama is likely to do is a constraint on what it is feasible for Merkel to do. But if

we want to know what it is feasible for Obama himself to do, or the estranged hockey father himself to do, then we make them the agents of our feasibility-for assessments. Maybe the results are informative; maybe they are uninformative. Perhaps knowing what people can do if they choose to is one consideration among many when it comes later to thinking about whether we can get them to do certain kinds of things. And perhaps it's just a useless bit of theoretical information. But that is an artifact of the concept. In the binary role, feasibility is about what options agents have. A theory is feasible if an agent has an option that could realize it. In the comparative role, feasibility is about the chance of an outcome, given an action of an agent's. Agents have many options that will go unchosen. Nonetheless, what feasibility *means* is what a person can do if they choose to – for the various reasons discussed in the last few chapters, that is how I have defined it.

Thus while it might seem appealing to try to limit feasibility assessments to those agents who we know take certain actions to be within their option set, we must resist trying to push feasibility and likelihood together in this way. The comparative feasibility tests do not ignore likelihood. In the first place, they conditionalize on something which is itself feasible, namely agents' choosing. In the second place, they always include a context, into which we can pack all the relevant background conditions and soft constraints, including whether the people around the agent make his action more or less likely to succeed. Conditionalizing upon an action is important because it lets us see the counterfactual worlds that are accessible from the spatiotemporal point we are interested in. Certain futures are open to the hockey father, even if they are not the future he will most likely end up in.

Cases closest to being purely theoretically interesting occur when a theory comes without restricted temporal scope, and for which we judge an outcome highly feasible because current agents have the option of coming together to constitute a collective agent which would have as an option an action that would with high probability produce an outcome realizing the theory. The limiting case is all current agents having the ability to constitute the collective 'everyone' / 'the whole world' / 'the global society' or some synonym of these. Because of course, if *everyone* chooses some action, and that action is reasonably likely to produce the collective outcome, then the outcome is highly feasible. Precisely the reason we judge so many political futures to be infeasible is that, given everything we know about people and politics, we just don't expect people to choose particular actions. But assessing feasibility-for 'everyone' is an excuse to conditionalize

upon everyone's choice of an action, and all that is left to the context are the environmental and physical constraints rather than the dynamic constraints posed by other people's psychologies. Is it feasible for the global society to solve global poverty if it tries? Yes, probably. Is it feasible for the global society to open borders and achieve global peace by allowing people to 'vote with their feet' for the kinds of societies they want to live in? Yes, probably. Is it feasible for the global society to achieve a perfectly egalitarian distribution of wealth and opportunity between all people? Yes, probably. These are strange answers to questions about 'what is feasible', admittedly. But they should not be taken as counterexamples to the formulation of feasibility in general. Perhaps it is of some interest to know what the global society could achieve if it chose to, despite knowing that it will almost certainly never (be in a position to) choose to, because current agents will almost certainly not choose the option available to them to *constitute* that agent (a precursor of its having the collective agency necessary to action choice). If it is of no interest, then we just shouldn't spend our time and energy asking what is feasible *for the global society*. Rather we can ask what is feasible for us – activists, policy-makers, citizens, democratic majorities; or for them – when 'they' are the political leaders we act as advisors for, or people who have expressed an interest in our analyses of what is feasible for them, and so on. The answers to questions about what is feasible for us, or for others who we can expect to choose the actions we conditionalize upon, might be more useful and more interesting with respect to making policy or arguing for non-ideal political theories, but the answers to questions about what is feasible for distant others are no less answers to questions about *what is feasible*, for that (although these considerations apply only to the general tests, not the agent-relative, because for the latter we can just say there *is* no such agent).

However, even though I have said we should resist trying to add an epistemic condition to making feasibility assessments, i.e. saying that the agent must know that an action is a potential action of hers, we might still want to add an epistemic condition with respect to those *using* feasibility tests. Some action and outcome pair is feasible only to the best of our knowledge. We should not draw conclusions in ignorance; they should be based on the relevant empirical evidence. These facts will often be determined contextually, made obvious by the kind of feasibility question we are asking. For example, if we are asking about whether a certain adaptation in human infants is feasible, the relevant body of empirical information is the biological literature on early human development. Or if we are asking about whether a certain infrastructural project

is feasible, the relevant empirical information will be on the state's budget, the availability of necessary resources, and so on. But we are fallible, and we will almost never be in a state of perfect information. For that reason there are epistemic limitations upon the *assessment* of feasibility claims. We will judge that a project is more feasible the more that it does not clash with soft constraints, to the best of our knowledge. That is an important caveat to remember, because it means that our judgements about what is feasible will change across time – without contradiction – as more information becomes available.

In summary, I have said that we should resist an epistemic constraint upon actions available to agents (from the agent's own perspective), but accept that our judgements about what actions are available to agents and what outcomes actions can bring about are reliant upon the best evidence, e.g. are epistemically constrained).

6.3 Applying feasibility: the normative element

The binary feasibility tests restrict the pool of recommendations, while the comparative feasibility tests allow us to rank them in order from more to less feasible. All of the recommendations in the ranked pool are potentially choice-worthy, depending on how they intersect with desirability considerations. If the second- highest-ranked alternative in terms of feasibility is much more desirable than the highest-ranked alternative, and their difference in feasibility is not enough to make the success of the second- highest-ranked unlikely, that might be a reason to choose it. By 'desirable state of affairs' I do not necessarily mean to refer to the sociology of desire, what people happen to desire, but rather to what is genuinely desirable, or desirable from a certain externally fixed perspective. There are two important questions we need to ask when we introduce the normative element as a partner to feasibility assessments in reaching all things considered political judgements. The first is, is the desirable state of affairs more desirable than the status quo? If so, there is a prima facie case for moving to it. The second is, what are the risks of pursuing each of the accessible pathways from the status quo to the desirable state of affairs? For each potential action of ours (e.g. the action intended to produce the desired outcome), we must ask what risks it entails. Any action has a chance of producing many different outcomes, and not all of these will be desirable. If an action has a fifty percent chance of realizing the desired outcome, and a fifty percent chance of creating a catastrophe, we should be reluctant to choose it (unless

the circumstances we are in are so unstable that they make it choice-worthy).

What is difficult is that there is no generally accepted body of moral knowledge that mirrors our body of scientific knowledge. Thus I cannot simply fill in the details, in any uncontroversial way, of what is desirable. In some cases, the moral principles and theories established at Stage One of theory construction (see discussion in Chapter 2) are what determine the normative considerations in Stage Two. Sometimes the normative considerations will be pre-determined by the ideology of those making the feasibility assessments (because often, when we choose to enquire into the feasibility of a certain set of options, that is because we already have a good idea of what we desire to do, and want to know whether it is feasible to do it). In other cases it might be positive morality that fills in the normative details, such as the current values of the general public, or the values of the judiciary, or the elected political leaders. What is politically desirable is an important question, and what is politically feasible is another. I have chosen in this thesis to be concerned with the latter. I do not wish to make any controversial claims in this section about the former. Suffice it to say that any all things considered judgement about what policy we should actually pursue will require input from both sources: we need to know what is desirable, we need to know what is feasible, and we probably need to be able to make tradeoffs between the two.

6.4 Decision Theory

In the last two sections I argued that in addition to feasibility considerations, epistemic limitations and normative considerations are also important when it comes to deciding which political recommendations we should actually try to bring about. I imagine that these considerations will combine in something like the following way. For any recommended outcome, e.g. ending global poverty, we'll have to think about all the ways we can possibly achieve it. For each of these ways, we'll have to rule out those that are infeasible in the binary sense. If all the ways a proposal can be realized are infeasible in the binary sense, then the desired outcome is itself infeasible in the binary sense. If all the ways of ending global poverty are infeasible in the binary sense, then ending global poverty is infeasible. If only some of the ways a proposal can be realized are infeasible in the binary sense, or if none of them are, we rank the remainder according to their feasibility in the comparative sense. Let's assume that it is more feasible to end global poverty via a system of taxes on luxury goods and travel than it is to end it by soliciting

voluntary donations from developed nations. Then we have to ask, for each of these actions, what their possible outcomes are. The desired outcome is one, but all actions come with risks of alternative possible outcomes, and as mentioned, not all of these will be desirable. For each of the possible outcomes of the given action, we have to think about how desirable it is, and the likelihood that the action will produce that outcome, given the best evidence we have.

These considerations can be formalized using decision theory. A basic formula borrowed from that literature goes something like this:

$$U(A) = \sum_o V(O) \times P(O \mid A)$$ [35]

Which is to say that the value of an action *A* is equal to the sum, across all possible outcomes *O* of that action, of the value of *O* multiplied by the probability of *O* given *A*. We can use this formula to calculate the values[36] of competing actions under consideration, such that the action with the expected subjective value no lower than any other should be preferred. This seems to have all the elements we have discussed so far, although notice that it is a tool for selecting which *actions* to pursue, not which *outcomes* to pursue, and the latter will often be politically more pertinent. We have a way of assigning a final value to actions, which allows us to compare and rank the actions under our consideration in terms of which is best, all things considered. We can complicate the formula slightly further to include the epistemic limitations mentioned in Section 6.2:

$$U(A) = \sum_o V(O) \times P(O \mid A \,\&\, E)$$

The formula now reads: the value of an action *A* is equal to the sum, across all

35 For simplicity, the formula here uses evidential decision theory with conditional probabilities. But it is worth pointing out that there is a great deal of debate between evidential and causal decision theorists. Causal and evidential decision theory *agree* on a large number of mainstream cases. They yield the same output, or recommend the same course of action, when added to the assumption that we should maximize rational utility. But there are some cases, far out in logical space, where the two come apart, and for those cases it is important that we figure out which of causal and evidential decision theory give the right answer. In my opinion, causal decision theory gives the right answer in e.g. Newcomb-type problems. I flag the issue only in the unlikely event that a case where the two theories come apart should arise in the course of our considerations. But for the remainder of the thesis I will just speak in terms of "decision theory" in general, and use the formula of evidential decision theory to preserve the simplicity of the calculations. On evidential decision theory see e.g. (Price, 1986, pp. 195–212), (Horgan, 1985, pp. 159–182), and on causal decision theory see e.g. (Joyce, 1999), (Lewis, 1973), (Sobel, 1994), and (Skyrms, 1980).

36 I mean 'value' to be used interchangeably with 'utility' or 'desirability' or even 'payoffs'. I use 'value' to avoid the connotations of 'utility' coming from economics, but I do not mean for that choice to imply a preference for evidential over causal decision theory.

possible outcomes *O* of that action, of the value of *O* multiplied by the probability of *O* given *A* and the available evidence (see Mellor, 2005). What is nice about borrowing from decision theory is that we can use some of the debates going on there to enrich our thinking about feasibility in practical reasoning. For example, we might ask whether groups like the state's agencies, political parties, social movements, all of whom might be subjects of a feasibility assessment, can properly be said to have beliefs; if not then the decision theory formula intended for individual agents may fail to apply. And even if groups can have beliefs, there's a problem in assigning beliefs about probability and value to them. If we take a majority decision, we risk aggregation problems (see e.g. List & Pettit, 2005; List & Deitrich, 2008); if we assign a "leader" whose opinion is to count for everything, we risk dictatorship; if we take empirical evidence as sufficient without judgment, we risk a standoff when experts disagree about plausible interpretations of the data. There is also the risk that value may be predetermined by a given ideology or school of thought. In that case practices of critical public deliberation may be helpful, but then we will have the problem that such deliberations may not yield converging views (more on collective beliefs and preferences in Chapter 8, and feasibility for collectives in Chapter 7).

The fact that I have said that feasibility can be formalized *within* decision theory should indicate that I do not think it can *replace* decision theory. Feasibility and desirability are both crucial to making an all things considered choice about which actions to select. Also, decision theory is not used as a tool to select which *outcomes* to pursue, although feasibility will often contribute to that decision. Although it is only a part of decision theory, feasibility can play two roles within it. First of all, it can simply give us more resources for making judgements about what is likely. The decision theoretic calculus requires us to figure out a conditional probability, the probability of an outcome given an action and the available evidence. One way we can do that is by appealing to the hard and soft constraints discussed in Chapters 4 & 5. An action has a zero probability of being done if there is no agent who has it in her option set; an outcome has a zero probability of being produced if there is no action that could bring it about. Whether an agent has an action in her option set, and whether an action can produce an outcome, are determined by the hard constraints, timeless, time-indexed, agent-relative, and context-sensitive. An outcome has a positive probability when the chance of an outcome *given* an action (available to some agent) is greater than zero. The chance of the outcome given the action determines the degree to which the outcome is

feasible. What chance actions have of producing outcomes is determined by the soft constraints bearing on a particular context.

Second of all, feasibility can combine with rough judgements of desirability and act as a decision making heuristic. Sometimes it is not possible to run the full decision theoretic calculus. And when it is not possible to do things decision theoretically, we might just use a rough and ready heuristic. We might ask 'how feasible is this outcome?' and 'how good would it be?' We try to get the highest score on both, making tradeoffs between the two as necessary.

An objection to feasibility in this role as a heuristic for decision making would be that it gets things wrong with respect to action choice, or choosing which outcomes to pursue. When we only care *how feasible* an outcome is, we ignore how *risky* it is. If an outcome has an 80% chance of being brought about, given an action, then we'll prefer it to an outcome that has a 70% chance of being brought about, given an action. But the remaining 20% in the former might be a risk of catastrophe, while the remaining 30% in the latter might be a chance of something much more desirable. The idea is that the heuristic can get things wrong because it just tells us to pursue the outcome that is most feasible and most desirable, which can turn out to be the wrong choice if the outcome comes with a high risk of serious harm. One way to come back at this objection is to say that it's part of *what's desirable* to think about whether to pursue a highly feasible outcome with a chance of serious risk. The other is just to concede that invariably, heuristics sometimes get things wrong, and so long as this one gets things right in the majority of cases, we are justified in using it.

I said in Chapter 4 that when an option is infeasible in the binary sense it is 'ruled out' of consideration for implementation. Feasibility assessments make a definite recommendation to policy-makers: do not pursue outcomes that are infeasible! But what if *pursuing* that outcome would produce more gains in a consequentialist calculus than pursuing an outcome that is feasible, in the binary sense, would do. Sure, we can't actually bring the outcome about. But if we can do better pursuing infeasible outcomes than feasible outcomes, oughtn't we to do that? The response to this is to distinguish campaign strategies from political decisions. Policy makers ought only to pursue outcomes that are feasible in the binary sense, and their desirability being equal, they ought to pursue outcomes that are more feasible over outcomes that are less feasible. But that is to say nothing about how pursuing those outcomes should be packaged. For example, imagine that policy-makers figure out that zero carbon emissions worldwide is

strictly infeasible. And they determine further that about the best trade-off between feasibility and desirability would come from pursuing the outcome of a two-thirds cut in carbon emissions from current levels. But, having decided that, imagine that a campaign for people to cut their emissions to zero would likely result in their cutting their emissions by two-thirds, and a campaign for people to cut their emissions by two-thirds would likely result in their cutting them by one-third. The former produces more gains. But that does not contradict the decision of the policy-makers. The outcome of cutting by two-thirds is feasible; one way to achieve it is via a campaign to cut to zero. The outcome of cutting to zero is infeasible; there are no agents with the actions available to them that could bring it about.

6.5 Counterfactuals: further epistemic limits on feasibility assessments

Claims that something is feasible are usually claims that certain kinds of actions *will* produce certain kinds of outcomes. Or, if they are about the past, they are usually claims that certain kinds of actions *would have* produced certain kinds of outcomes. These claims, respectively, are statements incorporating indicative conditionals ('if I do *x*, then *y* will happen'), and statements incorporating counterfactual conditionals ('if it were that *x*, then it would be that *y*'). But are we good at assessing the truth of counterfactual and indicative conditionals? If we are not, this whole project is in trouble.

Because claims about feasibility in political philosophy are almost exclusively about future states of affairs, I will talk in terms of the future indicative conditional and the future counterfactual conditional. Many people think the two are indistinguishable. Calling future-oriented conditionals 'counterfactuals' is slightly misleading, because of course one such counterfactual will turn out to be actual, and therefore not counterfactual at all. But because the future is open and we cannot know which of the possible alternative futures will be the actual one, I will speak as if they are all counterfactual.

Counterfactual reasoning is undoubtedly both widespread and useful. Several authors have shown that counterfactual reasoning is associated with better performance on future tasks (see e.g. Byrne, 2005; Kahneman & Tversky, 1982; Roese & Olson, 1993; 1995; and discussion in Williamson, 2007, p. 140). For example think of a child who puts her hand into the fireplace and gets burned. If she reasons 'if I hadn't put my hand in the fire, I wouldn't have been burned', she is much more likely to avoid putting her

hand in the fire in the future. Counterfactual reasoning can aid learning by trial and error. Or suppose the same child tidies her room without being asked, and is given a special dessert as a reward. And suppose she reasons 'if I hadn't tidied my room without being asked, I wouldn't have been given this dessert', which makes her more likely to do her chores unasked in the future. Counterfactual reasoning can help individuals to secure greater rewards. Unfortunately research on counterfactual reasoning is still in its infancy: 'as for the psychological study of the processes underlying our assessment of counterfactual conditionals, it remains in a surprisingly undeveloped state, as recent authors have complained' (Williamson, 2007, p. 142).

But given how widespread counterfactual reasoning is, we might expect people to be quite good at it. In that case, what's the problem? The problem is that reasoning about simple counterfactual conditionals like 'if I had let go of the helium balloon, it would have fbated up into the sky' is just *easier* than reasoning about complex counterfactual conditionals like 'if indigenous Maori had been more aggressive when the first colonists came to New Zealand, New Zealand would not now be an English colony', or 'if we can persuade people to support the 'say no to a bag' campaign, we'll be able to convince them later to support animal liberation and carbon reduction'. These kinds of counterfactuals are harder because it's hard to know whether English colonists, having had a bad first experience in New Zealand, would have backed off, or would rather have renewed their efforts with increased vigour, and it's hard to know what people's reasons are for supporting the 'say no to a bag' campaign. That's not to say that the evaluation of such conditionals is impossible, it's just to say that it's not easy. In the rest of the chapter I want to consider standard treatments of counterfactuals, to figure out whether and how we are supposed to be able to figure out when they are true.

6.5.1 Williamson

Timothy Williamson in *The Philosophy of Philosophy* (2007) is engaged in the project of defending knowledge of metaphysical modality, which he does by arguing that it is subsumed under our ordinary cognitive capacity to handle counterfactual conditionals. In the course of the discussion he argues that such conditionals are commonplace in our everyday lives, and that we have evolved mechanisms for figuring out when they are true and when they are not, mechanisms which can be extended to explain our capacity to distinguish conceivability from inconceivability (or metaphysical

possibility from metaphysical impossibility). Williamson's example involves a scenario in which there is a steep mountain, with stones embedded in the ice that are freed as the ice melts, and with a bush in the middle of the slope, and a lake at the bottom. He asks us to assess the truth of the counterfactual conditional 'if the bush had not been there, the rock would have ended in the lake' (Williamson, 2007, p. 142). He says that the truth of that conditional is quite obvious, and the reason we know that it is true is that various kinds of other facts allow us to figure it out.

What kinds of facts allow us to see that those conditionals are true? Probably we can figure out just by knowing the laws of nature what will happen in Williamson's scenario. If we know about gravity then we know the rock will roll downward; if there is a hedge in its way it will be stopped, if there is no hedge and it is otherwise unimpeded it will continue down the slope and end up in the lake. Knowing how certain kinds of objects behave in certain kinds of conditions gives us a lot of information. Moreover, we can appeal to past experience to say what might happen in a given scenario. If we have seen rocks roll down slopes before then we will know from experience that the bush is likely to stop the rock, and that in the absence of the bush the rock is likely to gather speed and splash into the lake below. Depending how *much* we know about the relevant kinds of facts, we will be able to give more or less precise information about what is likely to occur. For example, if the rock has an uneven surface, and we know about how those kinds of surfaces behave in interaction with a surface such as the slope has, then we will be able to predict the rock's trajectory down the hill, perhaps guessing with a high degree of confidence where exactly it will enter the lake.

Another means of figuring out the truth of a given counterfactual conditional is to 'do the experiment': free some rocks on an icy slope, and see what happens. Yet another is to allow straightforward reasoning to do the work of assessing certain counterfactuals. For example, we know that 'if twelve people had come to the party, more than eleven people would have come to the party' (Williamson, 2007, p. 143). Likewise simulation processes (the mental adoption of another's perspective) allow us to figure out what people are likely to do in certain circumstances, or what we ourselves would be likely to do in circumstances other than the ones we are currently in. Of course, our imagination is not limited to what will probably happen in certain circumstances. Williamson acknowledges that we have the imaginative ability to conceive of anything at all; we could imagine the rocks from his example flying up into the sky instead of rolling down the hill. Or in another example, a friend when asked a

simple favour could go mad and physically assault the asker, instead of congenially agreeing. But in general we don't imagine scenarios like the latter because 'the imaginative exercise is richly informed and disciplined by [...]our sense of what she [the friend] is like' (Williamson, 2007, p. 148).

Williamson notes that knowledge of some counterfactuals comes from expectations that are hardwired into us, like knowledge about the behaviour of fast-moving objects. It will have been evolutionarily advantageous to know about the behaviour of fast-moving objects, in order to avoid being hurt, or eaten. But other kinds of expectations are not hardwired, like those that have only become relevant in recent years, for example expectations about how things would have turned out if the outcomes of specific political elections had been different (Williamson, 2007, p. 150). His claim is that 'where our more sophisticated capacities to predict the future are reliable, so should be corresponding counterfactual judgments' (Williamson, 2007, p. 150).

The use of expectation-forming capacities to judge counterfactuals corresponds to the widespread picture of the semantic evaluation of counterfactual conditionals as "rolling back" history to shortly before the time of the antecedent, modifying its course by stipulating the truth of the antecedent and then rolling history forward again according to patterns of development as close as possible to the normal ones to test the truth of the consequent (Williamson, 2007, p. 150; cf. Lewis, 1979). Of course, if we needed to 'roll back' history to make a certain potentially counterfactual proposal about the future come out as true, that would be a good indication that the proposal should count as infeasible, because we cannot in fact change the past. It might be interesting to know that we *could have* avoided e.g. the human rights abuses of Nazi Germany, but that will do no work (except indirectly) in influencing what we choose to do now.

However, knowing that we can avoid human rights abuses if we take certain paths of action now and cannot if we take others *will* do work in influencing what we choose to do now. So our expectation-forming capacities must work even harder than they do on both Williamson's and Lewis's views (more on which below). We must 'roll forward' history to the time of many different antecedents, and see if their (or their various alternative) consequents are likely to come out as true. But predicting the future is difficult because of the many odd, surprising and unpredictable things that can happen. For example, many people in the early twentieth century predicted flying cars in the twenty first, but no one predicted the rapid rise of the internet, and there are countless such examples. Williamson thinks imaginative simulation is not always

necessary for evaluating counterfactuals, and not always sufficient; and it is not always the case that they can be evaluated anyway (Williamson, 2007, p. 152). But imaginative simulation is probably the most *distinctive* means of analysis. He also acknowledges that sometimes a counterfactual will be neutral between competing outcomes, e.g. 'if the coin had been tossed it would have come up heads' and 'if the coin had been tossed it would have come up tails'. A fair coin has a 0.5 chance of landing heads, and knowing that, we have no reason to prefer one of those counterfactuals to the other as true (Williamson, 2007, p. 154).

> Despite its discipline, our imaginative evaluation of counterfactual conditionals is manifestly fallible. We can easily misjudge their truth-values, through background ignorance or error, and distortions of judgment. But such fallibility is the common lot of human cognition. Our use of the imagination in evaluating counterfactuals is moderately reliable and practically indispensable. Rather than cave in to skepticism, we should admit that our methods sometimes yield knowledge of counterfactuals (Williamson, 2007, p. 155).

Williamson's conclusion is that while we are of course fallible, our evaluation of counterfactuals is reasonably reliable, and in any case, we could hardly do without it given the enormous practical advantages it confers. So on Williamson's view, we have good reason to be optimistic about our general abilities to evaluate counterfactuals.

6.5.2 Lewis

> The way the future is depends counterfactually on the way the present is. If the present were different, the future would be different; and there are counterfactual conditionals, many of them as unquestionably true as counterfactuals ever get, that tell us a good deal about how the future would be different if the present were different in various ways (Lewis, 1979, p. 455).

David Lewis too is optimistic about our ability to reason counterfactually, and that the propositions expressed by counterfactual conditionals can be true. His analysis of counterfactuals provides one of the seminal definitions in the contemporary literature:

> Roughly, a counterfactual is true if every world that makes the antecedent true without gratuitous departure from actuality is a world that also makes the consequent true. ... A counterfactual "If it were that A, then it would be that C" is (non-vacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C is false (Lewis, 1979, p. 464-465).

Take the conditional 'if I were to have pressured you, then you would have voted in the election'. On Lewis's analysis, this is true if for all the worlds in which I pressured you, those in which you also voted are closer to the actual world than those in which you failed to vote. If your voting would require a radical shift in personality (imagine you

have refused to ever vote in your life), then the world in which you feel pressured so you go and vote will be much less similar to the actual world than the world in which you feel pressured but nonetheless refuse to vote. Similarity is about the holding fixed of as much relevant information in the actual world as possible. But similarity is not a relation that is unchanging. Lewis says that different ways of weighting or prioritising the different kinds of similarities and differences are appropriate to different contexts. Some respects of similarity will obviously be irrelevant, like the ratios of vowels to consonants in the works of two writers, or whether two emeralds are both 'grue' (Lewis, 1979, p. 466). But others will obviously be relevant.

Lewis uses one particular objection to his analysis of counterfactuals as a means of coming up with a rough list of priorities for similarity. The objection is that a world without a nuclear holocaust is surely more similar to the actual world, all things considered, than a world with a nuclear holocaust. Yet 'if Nixon had pressed the button, there would have been a nuclear holocaust' is surely true (Lewis, 1979, p. 467, citing Fine, 1975, p. 452). One way to make it false is to allow miracles, i.e. violations of the laws of nature, so that for example Nixon might have pushed the button but somehow, miraculously, the signal failed to trigger the holocaust. Then that counterfactual, initially proposed by Kit Fine, is false, because the closest worlds to the actual world where Nixon pressed the button are nonetheless worlds where there was no nuclear holocaust. After some discussion, Lewis concludes that the priorities should be: (1) it is of primary importance to avoid large miracles; (2) it is of secondary importance to match worlds across time and space in terms of particular fact; (3) it is of tertiary importance to avoid even small miracles; and (4) it is of little or no importance to secure approximate similarity of particular fact (that's because small differences ramify outwards, so approximate similarity won't count for very much, for very long).

That is to say, Lewis does not take the strategy that would make the Nixon counterfactual come out as false. Because small differences ramify outwards, even the fact of having pressed the button without it triggering a nuclear holocaust will be sufficient to make the world different from this one. To 'clean up' all such traces would require a large miracle, but it is of the first importance to avoid those. To allow such traces would produce a world approximately but not exactly the same as this one, but it was of little or no importance to preserve approximation to particular fact, so we have no reason to prefer this approach. If there is a small miracle that can succeed in allowing the two worlds to match across time and space in matters of particular fact, then we

should allow such a miracle; but if there is not we should not. Lewis thinks there is not: 'I put it to you that it can't be done!' (Lewis, 1979, p. 473). The presumption then is that anything that can allow Nixon to press the button without also allowing a nuclear holocaust would require a large miracle, and so the closest possible world to the actual world where Nixon does press the button will in fact be one where there is a nuclear holocaust. So the counterfactual 'if Nixon had pressed the button, there would have been a nuclear holocaust', which is intuitively true, comes out true on Lewis's analysis after all, given a suitable understanding of the priorities of the similarity metric.

Some of the priorities of the similarity metric matter for our purposes more than others. In the first place, because we are in general concerned only with proposals about the future, about where we can get to *from here*, we won't need to worry about matching particular fact of futures. We can't worry about that, because we don't know the contents of the future. We can worry about matching the particular facts of history, of course; any world that would require a different history to make a future counterfactual come out as true will not be accessible (there will be no trajectory) from this point in the actual world. So the matching of worlds across time and space just tells us to hold our history fixed in thinking about what is feasible. Avoiding large miracles is important too – perhaps more important than it might at first seem. What it initially calls to mind is the changing of widespread facts about how the world is. But it might be something as simple as changing the character of one person, or making one unrealistic assumption about a population. Think about propositions like 'if the delegate for China doesn't stonewall at the next climate change meeting, we can probably realize the 2050 targets for carbon emissions reduction', or 'if we can get a few more poor people to vote in the next election, we will probably manage to swing back to a left-wing government'. These might seem like small changes to how the world actually is- China's delegate stonewalled at the Copenhagen meeting in 2009, and in general fewer poor people than rich people act on their right to vote. These facts might be fairly well entrenched by the preferences of major international players, and institutional and systemic factors about poverty in society. So an assumption requiring this 'small change' for a desired future to result might in fact constitute a large miracle on Lewis's understanding.

Of course there will be a real question about differentiating small miracles from large miracles, and evaluating all the facts that might need to be different for the world to be such that the antecedents of various counterfactual conditionals are true. And we will have to be able to decide which kinds of changes are optimistic but realistic, and

which are miraculous. For example, is the future in which more people vote in a current election than had voted in a prior election 'miraculous'? Or is it simply one of the many alternative possible futures that stands a good chance of becoming actual? We know that things can get better, because we've seen them get better. But we also know they can stagnate, and they can get worse. The key lies in assessing what we can realistically expect, and what we can realistically hope for. That is how we can figure out whether some counterfactual conditional under consideration is true or false.

6.5.3 Hawthorne

John Hawthorne writes in response to David Lewis's (1973) account of counterfactuals that it cannot account for a chancy universe. This is a problem because our best scientific theories tell us that our universe is chancy. Although an event with extremely low quantum- or statistical-mechanical chance is so small that probably no person in the history and future of the world will experience one firsthand, there is nonetheless a *chance* of such an event happening at any given point. Hawthorne uses the illustration of dropping a plate to make his point. Take the counterfactual conditional 'if I had dropped the plate, it would have fallen to the floor' (Hawthorne, 2005, p. 396). Given that there is a chance, however small, of an event occurring such that the plate, instead of dropping to the floor, instead flies off to the side, the counterfactual conditional will come out as false. But this is counterintuitive, because the example generalizes to mean that the propositions expressed by ordinary counterfactuals (which, as we have seen, are prevalent in everyday life) are also false. Hawthorne considers two solutions to this 'threat' to the truth of ordinary counterfactual conditionals. Instead of saying that the proposition expressed by 'if *p* had been the case, then *q* would have been the case' is true when all the closest *p* worlds are *q* worlds (following Lewis, 1973), Hawthorne proposes (a) modifying 'all' to 'most' such that the proposition expressed by the counterfactual is true when *most* of the closest *p* worlds are *q* worlds, or (b) taking 'closest' to rule out worlds in which a radically unlikely event actually occurs. He dismisses (a) and focuses his attention on (b).

Lewis's notion of a 'quasi-miracle', a remarkable event with low probability, is used as an example of the kind of thing that automatically makes a world more distant from the actual world. On this analysis, the counterfactual we began with, 'if I had dropped the plate, it would have fallen to the floor' comes out true, because in all the

closest worlds where I dropped the plate, it fell to the floor. The world in which the plate flew sideways is not a close world, because the fact that it contains a radically unlikely event affects its similarity. Hawthorne presents four problems for a development of Lewis's view along these lines.

One such problem is that there are some remarkable events such that although we might not expect to see them on any given occasion, we would expect to see them across some suitably long stretch of time. For example, although we would never expect a person flipping a fair coin one million times to produce a single series of all-heads flips, we would expect perhaps that for $2^{10,000,000}$ persons all flipping fair coins one million times each, one such flipper would produce the sequence 'all heads' (Hawthorne, 2005, p. 402). For any one of them, it seems like we want to be able to say that if they had flipped their coin it wouldn't have landed all heads, but if we say that for all of them, we get a counterfactual that is clearly false (this is just another example of the lottery paradox due to Henry Kyberg (1961, p. 197)).

Hawthorne also worries that the actual world contains lots of low probability remarkable events, i.e. quasi-miracles, and if the actual world contains lots of these then worlds that also contain them are no longer dissimilar in the way that was first thought to protect the truth of the ordinary counterfactual. He proposes to give up the Lewisian story altogether, instead opting for the idea that there is a *unique* closest world, such that the counterfactual conditional 'if it had been that p, then it would have been that q' is true if in the uniquely closest p world, q holds. But this requires knowing what holds at the uniquely closest p world. Hawthorne notices this:

> One might protest that there is a residual epistemological problem: how then can we know the truth of the counterfactual that if I had dropped the plate it would have fallen to the floor? Doesn't this require an utterly mysterious kind of modal insight? (Hawthorne, 2005, p. 404).

Hawthorne prefers that we 'give up on all neo-verificationist analyses of counterfactual discourse' (Hawthorne, 2005, p. 404). We have to, on his view, because the closeness relation between the actual world and the world where the antecedent of the counterfactual holds is unobservable. That means that the conditions that need to hold for a counterfactual to be true are unknowable, and thus the truth of ordinary counterfactuals is unknowable too. Notice how different this position is from Williamson's, which maintains that we have fallible but fairly reliable knowledge that many counterfactuals are true, and also Lewis's, which maintains that many counterfactuals are true, by way of his similarity metric. Hawthorne's worry is that all

ordinary counterfactuals are false, because their truth-value depends on something we cannot detect, namely the primitive closeness relation between worlds. 'If it had been that *x*, it would have been that *y*' is true if at the uniquely closest world where *x* holds, *y* holds. But this relation between *x* and *y* is unverifiable; there is no observational evidence that can count for or against its being present.[37]

To some extent, how radical we take Hawthorne's position to be depends on how seriously we take ordinary sceptical worries in epistemology. When we judge a counterfactual to be true by extrapolation from the laws of nature and our background knowledge (in the way Williamson suggests we do) we will very often happen to be right. There's no way of knowing whether the uniquely closest possible world is one in which the extremely unlikely event occurs, but all that this entails is that we'll mostly be right in our judgements, and every now and then in freak cases, be wrong. But Williamson already allowed that our judgements are fallible, and Lewis allowed that the similarity relation is not straightforward (in which case there will likely be error in judgement when applying it). In other areas of epistemology, many people think that the 'absolutely certain' kind of knowledge that would come from *knowing* we aren't in a world where a quantum event occurs is not necessary anyway. We can have knowledge that we have hands, even though there's some chance we are brains in vats, or even though in an extremely high-stakes bet we wouldn't be prepared to bet on the truth of that claim. So Hawthorne's position is radically sceptical only if we have that very demanding view of what counts for knowledge that a given counterfactual conditional is true. In any case, regardless of Hawthorne's own position, if political philosophers had to include with the claim '*x* is feasible' the caveat '...unless an event with extremely low quantum- or statistical-mechanical chance occurs' we can generally expect that to make no practical difference at all to their deliberations and subsequent recommendations about what we should do. In fact, given how cumbersome this caveat it, it would likely be dropped. It is not unreasonable to assume that people understand implicitly that when we make claims about what is likely to occur in the future, we are bracketing the possibility of freak events (of many different kinds) occurring.

6.6 Conclusion

In this chapter I have argued that when it comes to using the feasibility tests in

37 I am grateful to Wolfgang Schwarz for discussion on this point.

practice, we must also factor in important considerations about desirability, and must recognize that our assessments are limited by the best available evidence. These elements are essential to an all things considered judgement about which outcomes to pursue. I argued that we can use the formula of basic decision theory to capture these elements, and that feasibility in two further roles can supplement decision theory, first by constituting a judgement about what is conditionally probable, and second by acting as a heuristic when there's no time for the full decision theoretic calculus. In the last part of the chapter I noted that reasoning about feasibility is reasoning about counterfactuals, and I asked how good we are in general at that kind of reasoning. While there are reasons to be cautious, I argued against Hawthorn and with Williamson and Lewis that we can be reasonably confident in our ability to assess counterfactuals, and so in our ability to make assessments about what is politically feasible.

Chapter 7

Political Feasibility Revisited: Collective Feasibility

> Two neighbours may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each others mind; and each must perceive, that the immediate consequence of his failing in his part, is, the abandoning of the whole project. But 'tis very difficult, and indeed impossible, that a thousand persons shou'd agree in any such action; it being difficult for them to concert so complicated a design, and still more difficult for them to execute it; while each seeks a pretext to free himself of the trouble and expense, and wou'd lay the whole burden on others (Hume, 1978, p. 538).

7.1 Introduction

In Chapters 4 & 5 I gave a conceptual account of feasibility, in the former in its binary (feasible, or not) role, and in the latter in its graded (more or less feasible) role, and I put forward two tests of feasibility for each of these two roles. The binary tests are:

> **(General) Binary Feasibility Test:** The recommendations of some theory are feasible *iff* there exists an agent, or set of agents, within the designated period of time, who has available to her some action or set of actions that could bring the recommendations about.

> **(Agent-relative) Binary Feasibility Test**: The recommendations of some theory are feasible for an agent *iff* she has, within the designated period of time, an action or set of actions available to her that could bring the recommendations about.

For both of these, an agent has available to her some action if and only if her performing that action is not ruled out by any hard constraint. And an action could bring the recommendations about if and only if the action's producing the outcome is not ruled out by any hard constraint. The graded tests are:

> **(General) Graded Feasibility Test:** The degree to which the recommendations of a theory are feasible is established by the conditional probability of the recommendations being realized *given* an action (or set of actions), where the action is in the option set of some agent, and where the action is the most likely of any action available to any agent to bring the outcome about.

(Agent-relative) Graded Feasibility Test: The degree to which the recommendations of a theory are feasible for an agent is established by the conditional probability of the recommendations being realized *given* the action (or set of actions) most likely to bring the outcome about available to the agent.

In these tests, whether an action is in the option set of some agent is established by the agent-relative binary feasibility test, and the extent to which an action is likely to produce an outcome is established by soft constraints.

I said in Chapters 4 & 5 that the agents in the binary and comparative tests could be either individual or collective. Some of the examples I've discussed have involved collective agents, e.g. Carrot Mob and *Krawall & Remidemi* (Section 4.3), the New Zealand government (Section 4.4), and the apple pickers' union (Section 4.3), but many examples have involved individual agents, e.g. Jens-Christian and Weng Hong (Section 4.3 & 4.5), Clas and Astghik (Section 5.4.4), Leone and her flatmates (Section 5.6), Leon (Section 4.3.1), and Bill (Section 5.4.2). Agency extends in some cases to groups, such as policy-makers, activists, companies, and states. The feasibility tests are in terms of actions being in agents' option sets. But obviously it is easier to say what actions are in the option sets of individual agents than it is to say what actions are in the option sets of collective agents. The question of this chapter is, do any interesting problems or issues arise in thinking about the feasibility of *collectives'* action in particular? Notice that this is a distinct question from the problems or issues surrounding *collective action*, formalized in game theory. Game theory is about the problems that confront individuals in producing collective goods. But it models individual decisions, and action-choice (not collective decisions, or outcome selection). Collective action problems are often problems precisely because there *is* no collective, and coordination is not possible. The topics overlap, but their subject matter is distinct. What I will be concerned with in this chapter is (a) what a collective obligation means for members of the collective, because then it is possible to ask both whether the collective has the ability to fulfill its obligation, and whether the individuals have the ability to fulfill its (their) obligations, and (b) what it means to say that a collective has an ability, and where collective abilities come from. Answering these questions is crucial for the binary test in particular, because hard constraints function to rule an action out of an agent's option set. I want to know how that works in the case of collective agents.

Suppose that a theory of justice in transfer tells us that New Zealand is obliged to

provide redress to New Zealand Maori for historical injustices. We might want to know whether it is feasible that New Zealand fulfill the alleged obligation. If not, there cannot really be an obligation, although an evaluative claim ('it would be good if...') might remain. The agent-relative binary feasibility test would say that the obligation stands if New Zealand has some action in its option set that could bring about the outcome of redress being provided to New Zealand Maori for historical injustices.

There are two things to think about. The first is, how do we establish that a collective agent has some action as an option? That is the same as thinking about what abilities, or capacities, or powers, collective agents have, and in virtue of what they have them (which is to return to the role of feasibility mentioned in Chapter 1 and discussed in Section 4.5, of identifying 'powers'). The second is, how should we understand the relation between collective obligation, and the obligations of the members of collectives? Collective obligations fall upon collectives. Can they be infeasible if they are infeasible for the *individual members*? Or only if they are infeasible for *the group*? What does this distinction actually amount to?

Various topics approach these questions, such as the literature on collective action, discussions of shared intentions and shared cooperative activity, and discussions of collective *responsibility*, but none ask quite the same thing. Insofar as the discussion about collective responsibility is divided between methodological individualists, who defend the idea that all group responsibility reduces to individual responsibility, and methodological holists, who defend the idea that groups can be responsible independently of their members, I side with the latter. Answering the question of how we should understand obligation and ability for groups will be relevant to determining blameworthiness for both collectives and individuals when they are members of collectives. Feasibility considerations play a role in establishing blameworthiness because if an agent ought to have done φ, and it was feasible for her to have done φ, then she will be blameworthy if she failed to.

At this point I want to introduce a case, which I will use throughout the chapter to discuss both ability and obligation in the case of collectives. I will deal with obligation first of all, because only when we know what a collective obligation *means* in the case of individuals, can we ask whether it is feasible for the group, and its individual members, to realize the obligation, i.e. to ask whether some action producing the collective outcome is feasible for the collective, and its individual members. There will be some overlap between the discussion of obligations and the discussion of abilities. That's

because if there is an obligation at all, it has to be feasible; if it's not feasible then there's no obligation. This means a good account of the relation between collective and individual obligations and abilities requires charity. It would be all too easy to explain the relation in a way that meant the obligations were practically always ruled out as infeasible.

7.2 Obligation in the case of collectives

The case I will work with involves a collective, a *company*. It is a furniture removal company, constituted by four members, Kewa, Tom, Mark, and Jonno. The company has a business agreement stating that each member is an equal shareholder in its earnings, and an equal participant in its undertakings. The explicit agreement is cemented by the agreement implicit in the fact that the members have been friends for a long time. It happens that one day, in the process of shifting furniture out of an apartment, the boys stumble in the stairwell, dropping the heavy piano they are carrying and in the process hurting a small child. The child is trapped, and if she is not soon released, the weight of the piano will kill her.

Many, perhaps all, moral theories would require that the company lift the piano off the child. It is imperative to provide redress for harms you have caused; it is imperative to provide aid wherever there is suffering and it wouldn't cost you too much to do so; there is more good in the world in which the piano is lifted; lifting the piano is what a good person would do.

Notice that I supposed the moral theories would require that the *company* lift the piano. This might seem a little odd, because while it was the company engaged in removing furniture from the apartment, surely the situation can be described both ways: the company dropped the piano, or, Kewa, Tom, Mark and Jonno dropped the piano. We might even want to point the finger, and say that they only dropped the piano because Jonno stumbled. There are two things to say about this. The first is that I supposed the company dropped the piano because I am interested in how collective obligations, in this case the obligation upon the company to lift the piano, devolve to their members. The second is that while some people might prefer to resist collective obligations, I suspect one reason is that they don't think collectives are the kinds of things that can realize obligations. But that is precisely the question at issue. The aim of this chapter is to tell a convincing story about when a collective has the ability to realize

its obligation. If I can do that, one good reason to resist collective obligations disappears.

So assume for the sake of argument that it is the company which is obliged to lift the piano off the child. Now the question is, what does that collective obligation *mean* for the four men who make up that collective?

7.2.1 Collective obligations without member obligations

One possibility is that it doesn't mean anything. We might just say that the group has an obligation, and the individuals don't. Frank Jackson & Robert Pargetter suggest at one point in their (1986) that we should deny distribution over conjunction in the case of 'oughts' (see also discussion in Lewis, 1973, pp. 79-80). The cases they talk about involve more and less ideal circumstances, but we can apply the same discussion to groups. The structure of their idea is as follows. It ought to be that *A* & *B*. If 'ought' distributes over conjunction then it follows that it ought to be that *A*, and it ought to be that *B*. But now imagine, given that it *won't* be the case that *A*, it ought not to be the case that *B*. That creates a contradiction: it ought to be the case that *B*, and it ought not to be the case that *B*. Or, to put this in case form, it ought to be that the procrastinating professor accept the task of reviewing a book, and actually review the book. So, if 'ought' distributes over conjunction, it ought to be that the professor accept the review. But given that he won't review the book, it ought not be that he accepts it. The world in which he accepts then doesn't review it is worse than the world in which, knowing that he won't review it, he rejects the assignment, and the journal editor sends it to someone else. This creates a contradiction: he ought to accept, and he ought not to accept.

The solution Jackson & Pargetter suggest is a partitioning of option sets. Relative to the set of options including the actions 'accept and review' and 'accept and don't review', the Professor should accept and review; relative to the set of options including the actions 'accept and don't review' and 'reject', the Professor should reject. What he should do depends on what options we consider to be available to him (Jackson & Pargetter, 1986, p. 254). That looks like a neat solution, and at least in ordinary language it avoids the contradiction. It is not that all things considered the professor ought to accept and review, and also reject. It's only that relative to one set of options he ought to accept and review, and relative to another he ought to reject. (It's not clear that this is such a neat solution for the logicians, however, because it's not clear that deontic logic

can handle an obligation being *relative* to a particular set of options).[38]

One reason to reject Jackson & Pargetter's solution is that the procrastinating professor does not have more than one option set available to him. He has just one, and what he ought to do depends on what is in it. The fact that he won't write the review does not suffice to limit his option set in a way that changes what he ought to do, i.e. it does not make another option set the right one to appeal to. What he ought to do is the best of what he can do. If the best action in his option set is accepting the assignment and writing the review, then that is what he should do.

But it's not clear whether we can say the same thing for all pairs of ideal and non-ideal obligation statements. Assume that it's bad for people to carry weapons around. And now suppose the following is true: it ought to be that women are never assaulted, and do not carry defensive weapons around. If 'ought' distributes over conjunction then it ought to be that women do not carry defensive weapons around. But now suppose that the following is also true: given that women are assaulted, they ought to carry defensive weapons around. That creates a contradiction: it ought to be that women carry defensive weapons around, and it ought to be that they do not carry defensive weapons around. When we think about the way the world ought ideally to be, in comparison to the way it ought to be given some of the ways it actually is, plenty of contradictions with the same format will arise. And I'm not convinced that it will always work to tell a story about relativized option sets. That is one reason to think that Jackson & Pargetter's first inclination, to deny distribution over conjunction for 'oughts', was correct.

Their discussion was about different things that a person allegedly ought to do. But we can apply it to different parts of things that a collective ought to do. Let's parse the statement 'the company is obliged to lift the piano off the child' as 'Kewa and Tom and Mark and Jonno are obliged to lift the piano off the trapped child'. If 'ought' does distribute over conjunction, then it will be true that Kewa is obliged to lift the piano off the trapped child. But supposing neither Tom, nor Mark, nor Jonno will help him, and supposing he can't lift the piano alone, he can't be obliged to lift it. That is because he cannot be obliged to do what he cannot do. Distributing 'ought' over conjunction seems

38 To be more precise, cases like the procrastinating professor and the samaritan paradox show that deontic logic cannot be monadic. But David Lewis, for example, has suggested moving to dyadic deontic logic instead of admitting that obligation cannot be modeled within standard modal logic (see e.g. Lewis, 2000, pp. 5-19). In any case, the issue of whether ought distributes over conjunction in ordinary language (whether 'ought *a* & *b*' means 'ought *a* & ought b') is a separate issue from how the logicians choose to handle ought formally.

to get things wrong in that case. That is a reason to resist distributing. And this paves the way for collective obligations existing when obligations upon members of the relevant collective do not. It might be true that the company ought to lift the piano, and not true that Kewa ought to.

In any case, I don't have to accept that ought fails to distribute over conjunction as a way to distinguish collective obligation from individual obligation (or reconcile collective obligation with individual obligation) if I can tell a good story about *how* collective obligation distributes from the collective to its members.

7.2.2 Distribution without change

One way the collective obligation obviously *doesn't* distribute is in the same form. The company is obliged to lift the piano off the small child, but the piano is too heavy for Kewa to lift alone. So, being charitable, it can't be that the collective obligation distributes to each member such that he is obliged to *lift the piano off the child*. The distribution must be more sophisticated.[39]

7.2.3 Distribution into shares

The more plausible, and familiar, suggestion is that the members are obliged to *do a part*, or *take a share* in doing what the group is obliged to do. We know that the company has four members. Assuming that it takes all four to lift the piano, a distribution into shares of the company's obligation would obligate each member to take one quarter of the weight of the piano each. The members' obligations *add up* to the group obligation.

Of course I have to be careful. I just said that the collective obligation to lift the piano falls upon the members as an obligation to lift one quarter of the (weight of the) piano. But distribution into shares doesn't mean distribution into *equal* shares. This is clear when we consider that one member of the company might be a lot bigger than another, or one a lot smaller. It is plausible that obligation is relative to capacity. If Kewa

39 Actually, there might be a few cases where this is what we should say, e.g. overdetermination cases where any one person could perfectly fulfill the group's obligation by acting unilaterally. For example, imagine that the group of philosophers attending the Tuesday seminar is obliged to provide the speaker with a glass of water. Any one attendee might fulfill the group's obligation by getting the speaker a glass of water. But these kinds of cases will be relatively rare; I shall concentrate on cases where more than one member has to act for the collective outcome to be produced.

is a lot stronger than Jonno, then Kewa should probably shoulder more of the weight of the piano than Jonno. This is just as we think that the poor are not obliged to donate as much to charity as the rich. So let's say that in a distribution into shares, the relative size of the shares is relative to the capacity of the group members. On this story I am supposing that an obligation upon a collective, in this case a furniture removal company, translates into a capacity-relative obligation upon the members of the collective, in this case Kewa, Tom, Mark and Jonno, to do a *part* of what must be done.

But now imagine that Kewa believes, with a high degree of confidence, that the others in the company are squeamish about small children suffering and will likely flee the scene upon realizing a child has been trapped by the piano. What is Kewa's obligation in that situation? Knowing that there is nothing he can do to shift the piano by himself, is he under any kind of obligation at all? (Remember we are concerned with his obligations as a member of the group, not the obligations that bear on him directly as an individual). While he might have all sorts of secondary obligations, for example to go and find other people to help him, I think he cannot have an obligation to lift a capacity-relative share of the piano.

It might be tempting to think that that is because he cannot take a capacity-relative share, in light of the fact that the others won't. That is to suggest that the obligation exists prima facie but dissolves as a matter of his inability. But we should resist this explanation. Rather I would insist that Kewa did not have a categorical obligation to take a capacity-relative share in lifting the piano in the first place. What his obligation was, as a member of the company, was to take a capacity-relative share in lifting the piano *given* a *belief* that the other members would do the same.[40] In the literature concerned with providing the necessary conditions for group action, theorists include a condition of common or mutual belief along the same lines. For example, Michael Bratman includes 'it is common knowledge between us that...' (Bratman, 1992, p. 338), Philip Pettit and David Schweikard include 'each believe in common that...' (Pettit & Schweikard, 2006, pp. 21-24) , and Raimo Tuomela includes that an agent 'believes that there is (or will be) a mutual belief among the participating members...' (Tuomela, 1991, p. 263).

40 Some might prefer to say that his obligation is conditional upon what the others *do* rather than what he *believes* they will do, and that his beliefs only provide an excuse. I resist this formulation here because I think obligation must remain practical: a person cannot have an obligation to do something she (reasonably) doesn't know about, e.g. an agent cannot have an obligation to rescue a child drowning outside in his swimming pool, if he had no way of knowing there was a child outside and anywhere near his swimming pool unsupervised.

7.2.4 Beliefs triggering obligations

If collective obligations distribute to members as obligations to do a part of a collective action, given a belief about what the others will do, then the furniture company's obligation to lift the piano and free the trapped child distributes to Kewa, Tom, Mark and Jonno as an obligation upon each to do a part of lifting the piano, conditional upon a belief that the others will do a part too. This story seems to get things roughly right in the piano case. Kewa doesn't believe that the others will do their parts, so he isn't obliged to lift a share of the piano. That is good, because his trying to lift it alone would be futile. Rather it is better that he fulfill a secondary obligation, perhaps looking for others to help him.

But let's try to get precise about what a belief-dependent obligation must look like. All I have said so far is that Kewa is obliged to do his part only if he believes that the others will do theirs. That is because for the particular task involved in the case, all the members are needed for the collective action to be produced. The first question to ask is, what kind of belief is sufficient? The second question to ask is, what exactly is the logical structure of the obligation?

In response to the first question, we shouldn't allow just any old belief. What if Kewa stubbornly refuses to believe the others will do their shares, despite strong evidence to the contrary? Imagine that Tom, Mark, and Jonno immediately upon dropping the piano bend down to pick it up again, and they simply pause at that point waiting for Kewa to take his corner. If Kewa fails to do his share at this point, we should surely say it's because he fails to realize his obligation, not because his obligation was only to do his share *given* a belief that others would do their shares, and he didn't have that belief. So we should add that the belief must be *reasonable*, which is a place-holder for the idea that it should be sensitive to the available evidence.

In response to the second question, I have to be careful how I formulate the belief-dependence of the obligation. Kewa has a conditional obligation. If he has a reasonable belief that others will do their shares, then he must do his share. 'If *a* then *b*' is logically equivalent to 'not-*a* or *b*'. So Kewa is obliged *either* to do his share, *or* not to have the reasonable belief that others will do their share. One way to fail to have this belief is to have an *unreasonable* belief that others will do their share. But having that belief seems like an undesirable way for him to be able to fulfill his obligation. So we have to formulate the belief-dependence in a way that doesn't allow such escape routes.

One way to do that is to formulate the conditional obligation negatively. It is obligatory that *unless* Kewa reasonably believe that the others won't do their shares, he does his share. If he does not have the reasonable belief that the others will defect (not do their shares), then he must do his share. 'If not-*a* then *b*' is logically equivalent to '*a* or *b*'. If a person does not have the reasonable belief then he must do his share; he must *either* have the reasonable belief, *or* do his share.

7.2.5 Four types of collective action scenario

So far I have only been considering the case in which there are four members of a group, and all must take some part of the collective action if the collective action is to be produced. And I have suggested that the obligation upon the collective to produce that action translates into an obligation upon each member to do a share of the action, conditional upon a reasonable belief that the others will also do a share. But there are others kinds of cases.

For example, what if the company has *eight* members, and the piano can be lifted by four? What if more than four people would make lifting even easier? Or on the other hand, what if more than four people would make lifting much more difficult, by getting in each other's way? What if one person couldn't lift the piano alone, but could push it enough that the child would suffer a little bit less? And what if for every member who pushed at it, the child would suffer less and less?

Which of these kinds of cases we're in matters a lot for what story we tell about how collective obligation distributes to members. There are four basic types of cases. The first I call a *joint necessity* case. The piano case is that type of case. For the collective outcome to be produced, it is necessary that every member of the collective act. That is why any member's action is conditional upon the others' action: no member can realize the *collective* outcome alone.

The case in which every member of the company can make things a bit better for the child by pushing at the piano, I shall call an *incremental good* case. The more a member contributes, the better; and the more members that contribute, the better. These are the easiest cases to deal with, because the collective obligation distributes in a way that's categorical. Each member of the company should do a share of what the company is obliged to do, regardless of his beliefs about what everyone else will do,

because the more that do a share (and the greater the share), the better.[41]

There are two other kinds of cases. One I shall call a *threshold good* case, and the other a case of *threshold good with harm*. When there are eight members of the group, the fact that it takes only four to lift the piano makes a threshold good case. Any four members taking a share of the action is sufficient for meeting the threshold and producing the collective action, so there are a further four members who aren't strictly required to do anything. (Although, the fact that more members helping would make things easier might be a reason to make a further distinction between the cases, to allow a stronger obligation).

When there are eight members of the group, and any more than four trying to lift the piano would be more of a hindrance than a help, we have a threshold good with harm case. In those kinds of cases the collective outcome is produced when the threshold is met but not exceeded.

The only case that doesn't require the distributed obligation to be conditional upon belief is the incremental good case. I would suggest that the distribution in that case is as follows:

(1) **Incremental good**. When a collective has an obligation to φ, every individual member of the collective has an obligation to take a capacity-relative share in fulfilling the obligation.

Otherwise, we need to conditionalize on beliefs, and in slightly different ways. I suggest the following:

(2) **Joint necessity**. When a collective has an obligation to φ, every individual member of the collective has an obligation to take a capacity-relative share in fulfilling the obligation, unless she has the reasonable belief that at least one other member of the collective will not take a capacity-relative share in fulfilling the obligation.

41 I am assuming here that the collective good is fixed, and the individual contributions are incremental advances towards it. The situation is different when the collective good is itself an incremental good. For example, there is presumably a collective obligation upon Australia to lower its carbon emissions, and the more it can lower them the better. In that case it doesn't make sense to talk about an individual Australian's 'share'. It is not that there is some fixed outcome that is divided between the number of members in the group, so that when each does his share (or more) it is good, and when more people do their share, it is good. It can't be, because there's no fixed outcome to divide up. So in cases where the collective good is itself incremental, it makes more sense to say that the distributed obligations are capacity-relative contributions to the collective *pursuing* the desired (or a desirable) outcome.

(3) **Threshold good**. When a collective has an obligation to φ, every individual member of the collective has an obligation to take a capacity-relative share in fulfilling the obligation, unless she has the reasonable belief that sufficiently many other members of the collective will take a capacity-relative share such that the collective obligation will be fulfilled.

Sometimes cases of this third type will be such that all that matters is the threshold being met. But in other cases, it might ease the burden on the others, or make things easier, if the same burden is shared between more members (this is not to be confused with the case where *it would be better if* more members contributed, which would be an incremental good case). For example, it takes four people to lift the piano. But imagine that a passer-by sees the predicament that the members of the furniture company are in, and sees that helping them would ease the burden on each, even though it is not strictly necessary, and it would produce the same outcome as when the four members lift the piano alone. Nonetheless, we might still want to say that the passer-by should contribute. In those types of cases we might want to add a caveat to (3): '...and it would not lessen the burden on those taking a capacity-relative share for her to contribute in addition').

(4) **Threshold good with harm**. When a collective has an obligation to φ, every individual member of the collective has an obligation to take a capacity-relative share in fulfilling the obligation, unless she has a reasonable belief that sufficiently many other members of the collective will take a capacity-relative share in fulfilling the obligation so that her own contribution would be detrimental to the collective obligation being fulfilled.

Just to clarify, the suggestion in (2) – (4) is that an individual member of a collective has an obligation to a conditional. It is obligatory that, unless she has the relevant belief, she contribute. This is different from saying that if she has the relevant belief, then she is obliged to contribute. The obligation ranges over the conditional, not just its consequent. In (3), the idea is that the agent contribute unless she's sure enough others will; in (4), the suggestion is that the agent contribute unless she thinks enough others will that her own contribution would be harmful. The reason the obligation has to distribute conditional upon beliefs in these kinds of cases is that there's no action that

every member of the collective must actually perform. Some must act, but others must refrain from acting (e.g. to avoid harm). Members of collectives can fulfill the individual obligations that devolve to them by being sensitive to what others will do.

I said in 7.2.4 that it is not sufficient that an agent merely have a belief about what the others will do, she must have a *reasonable* belief. A good test of the strength of this modifier is conspiracy. Imagine that the members of a collective conspire to free themselves of their obligation to do a part in the collective action by agreeing that they will each not do their parts. In virtue of agreeing to this conspiracy, each member of the group comes to have very good evidence for the belief that the others will not contribute. Each member has a belief that respects the available evidence, so the belief counts as reasonable. According to (2) - (4) the members are then free of any obligation to do a part in producing the collective action. But what is really going wrong in the conspiracy case? Surely members are blameworthy for conspiring to escape their obligations. But insofar as they had conspired, each is surely not blameworthy for not doing a part in producing the collective action, which she would believe with high confidence to be futile. I think the conspiracy case shows that (2) – (4) give the right answer.

Cases will vary with respect to uncertainty. I have assumed synchronic decision-making, where a member of a collective must decide on the basis of the evidence available to her whether the others will do their shares, and thus whether she will do hers. But not all cases are like this. Some involve diachronic decision-making, where the agent sees *that* others are contributing and decides in light of that fact that she will also contribute. Others allow communication and coordination, so that members can decide together how they will act, and who will take which burden (not all cases are like the piano case, where members do the same thing, i.e. take some of the weight of the piano. In some cases, members will have to perform quite different actions from one another. Think of the various parts involved in a large engineering company constructing a new nickel mine, for example). In situations where a lot of information is available about others' intentions, it will be easier to produce the collective action; in situations with much less information, it will be more difficult (more on the likelihood of collective action in Chapter 8). But this chapter is not about when collective action is easier or harder to produce. It is about what collective obligation means for members of the collective (and what collective ability means in general). I have suggested that a collective obligation will mean one of (1) – (4) above.

In summary, notice what has happened to the obligation in the original case. The furniture removal company was obliged to lift the piano off the trapped child. I asked what that collective obligation meant for the members of the collective, Kewa, Tom, Mark and Jonno. After considering several possibilities, I settled on a distribution of collective obligation to members where each is obliged to take a capacity-relative share in fulfilling the collective obligation, unless he has a reasonable belief that others will not do their share. Members' obligations to act are conditional upon their beliefs in all types of cases with the exception of the incremental good case. The piano case as I have set it up is a joint necessity case. In that type of case, if any member has the reasonable belief that at least one other will not do his share, and does not act *for that reason*, then the member has *fulfilled his obligation*. This is important. The group fulfills its obligation by *doing what it is obliged to do*, in this case actually lifting the piano off the child. And the members of the group fulfill their obligations by doing what they're obliged to do. But what they're obliged to do is *take a part in lifting the piano off the child unless they have the reasonable belief that at least one other will not*. If they have that belief, then they don't have to *do* anything. In Section 7.4 I want to consider the implications of this. But first, let's look at how collective *ability* works.

7.3 Ability for collective agents

It is hard enough to say when an individual has some action in her option set. It is even harder to say when a collective has. How do we establish, for example, whether it's true or false that the German military in the time of Hitler had the option of overthrowing Hitler in a military coup? In this section, I want to try to figure out how we should think about group ability, and whether group and member ability come apart.

7.3.1 The German military

One might reasonably claim that the German military during Hitler's reign had the *ability* to overthrow Hitler. The military was physically close to him in a way that few other groups were (compare with the students, the blue-collar workers, and so on). It had plenty of weapons, and strategic training. If any group was to succeed in taking down Hitler, it must have been the military. This is a reasonable claim, but how do we show that it is true? I want to suggest that group ability is determined by group members'

abilities. With obligation, the distribution was downwards. Groups have obligations which divide into parts as they bear on the parts of the group, the members. But with abilities, I think, the distribution is upwards. Members have abilities, and these are aggregated to determine a collective ability.

So if we want to figure out whether the German military had in its option set an action that could have resulted in a successful military coup against Hitler, we look at whether the individual soldiers had in their option sets actions that could have aggregated to form a successful military coup against Hitler. Did they? A coup requires intense strategizing and planning. It needs leaders, and it needs supporters. But think about the conditions under which individual soldiers in the German military were operating. Loyalty to Hitler was extremely fierce. The penalty for treason was severe – in all likelihood death. There were spies and informants everywhere. This means that no soldier could have (without high risk of death) started planning and strategizing in the way required to initiate a successful coup. If you don't know who you can trust, and the chances are that you can't trust *a lot of people*, the risks of trusting anyone would be too high. Thus closer inspection of the reasonable claim that the military could have overthrown Hitler reveals that it is probably false. The military had the ability to overthrow Hitler if the soldiers making up the military each had the ability to do their parts in overthrowing Hitler. But they didn't have the ability; the conditions prevented it.

The answer I have just given depends upon a certain understanding of what suffices for inability. When discussing the hard constraints that rule an action out of the option set of an individual, I did not include risks. It is one thing to say that an action *cannot* be done by an agent, it is another to say that it is one she *should not* do, perhaps because it would be foolhardy. But we are thinking now about what kinds of things suffice for collective inability. We could tell the same story, and say that a collective action is not ruled out if the parts of it are not ruled out for any member of the collective; and the parts of it are not ruled out for any member of the collective if the member has in her option set an action that *could* produce her doing a part. That would mean only the soldiers' forcible prevention from doing their parts would suffice to genuine collective inability. But that seems much too strong. The soldiers are prevented for all practical purposes. If one tried to begin planning the coup, he would soon enough confide in an informant, and the price of that would be death. This is true for any soldier. Most people would say the soldiers didn't really have the option of planning a coup, even though it is true that there's something they could have, very recklessly, done.

If that slightly weaker understanding of the inability of individuals constituting collectives is the right one, then despite the plausibility of the claim that the military could have overthrown Hitler, the inability of the individual soldiers to do their parts tells us that it is not true. The soldiers were unable to communicate in the way that would have been required to plan a military coup. They did not have the constituent parts of the military's ability to pull off a coup. If the members didn't have the abilities that were constituent parts of the collective's ability, then the collective didn't have the ability. And if the collective didn't have the ability, then it can't have had the obligation.

I should enter a caution at this point. If we *mistakenly* decide that the collective has the ability, and then we distribute the collective obligation to members of the collective, it shouldn't be surprising that individuals have the abilities to fulfill their distributed obligations, despite lacking the ability to do a constituent part of the collective action. The military is not able to overthrow Hitler. But if we mistakenly decide that it is, we will say that the soldiers have an obligation to do a part of what is required to overthrow Hitler, unless they believe the others won't (or sufficiently many others will, or their own contribution would be detrimental, or if overthrowing Hitler were an incremental good situation). The soldiers have the ability to fulfill *that* obligation, because it requires them either to contribute (which we have already established they cannot do) *or* to have the right kind of belief. They have the ability to have the right kind of belief, so they are in a position to fulfill their obligations.

7.3.2 The furniture removal company (again)

Now let's think about a situation in which a collective does have the ability to fulfill its obligation. Presumably the furniture removal company introduced above has the ability to fulfill its obligation. Its obligation is to lift the piano off the trapped child. We can infer that it has the ability to do so from the fact that it did so: it was lifting the piano before it dropped it. But even without this inference, we can determine that it has that ability. How? We look at what needs to be done, and then we ask whether the members of the group each have the ability to do the relevant parts of what needs to be done. In the company's case, this is straightforward, because the members all have to do something, and they all have to do the *same* thing. Kewa, Tom, Mark and Jonno each have the ability to take a capacity-relative share (roughly equivalent to a quarter of the weight) of the piano. If they each did this, the result would be the piano being lifted.

If Kewa, Tom, Mark and Jonno were all extraordinarily weak, so that it was beyond them to lift the piano (setting aside the problem of how they could have come to drop it) then I'd have to say that the company lacked the ability to lift the piano. Or if the company had only two members, and we were to hold fixed that it would take four to lift it, I'd have to say the same thing.

Groups have abilities because members have abilities. But notice that this doesn't work in the other direction. Individuals don't have abilities because groups have them, and groups don't automatically have abilities just because their members have them. The fact that Germany had the ability to beat England in the 2010 World Cup doesn't give Bastian Schweinsteiger the ability to beat England. Rather it is because Schweinsteiger and his teammates had the abilities they had that Germany had the ability to beat England. And the fact that Kewa has the ability to do a triple somersault is not a reason to say that the *furniture removal company* has the ability to do a triple somersault.

7.4 Asymmetry in blameworthiness

In this chapter I have tried to give an account of how collective obligations distribute in different kinds of cases to the members of collectives, and when we should say that a collective has the ability to fulfill its alleged obligations.

What is interesting is that collective obligation, and collective ability, do not come apart from individual members' obligations, and individual members' abilities. If a collective has an obligation, there's a story we can tell about what the obligations upon its members are. When a group has an obligation, the members have an obligation, in whatever modified form is appropriate. And if the collective has the ability to fulfill its obligation, there's a story we can tell about the members' abilities to do the constituent parts of fulfilling the obligation. When the members are able, the group is able.

What *does* come apart is blameworthiness. And the fact that a group can be blameworthy while an individual member is not might explain why people sometimes seem to think that group and member obligation, or group and member ability, also come apart.

How does blameworthiness come apart between groups and group members? I can demonstrate this by returning to the piano case. I said that the company has the ability to lift the piano, because the four members who make up the company each have

the ability to do the relevant parts of lifting the piano, namely taking roughly one quarter of the weight, a bit more or less depending on their size and strength. And I said the company is obliged to lift it, because it dropped it on the child in the first place, and probably for other reasons. So there is no inability of the kind sufficient to reject the obligation; binary feasibility constraints do not do any work there. Thus if the company fails to lift the piano, *it is blameworthy*. It is obliged to lift the piano, and it has the ability to lift the piano, so if it doesn't, then it fails to do what it is obliged to do.

But that is not necessarily true for the members of the group. The collective obligation distributes to them in the form that they are obliged to take a capacity-relative share in lifting the piano unless they have a reasonable belief that the others will not take a share. So if any member has good reason to believe that at least one of the others will fail to take a share (because remember they are jointly necessary), then he will not be obliged to take his share. The collective action can fail without the individuals being all to blame. In fact, the collective action can fail without *any* individual being to blame. So long as the beliefs are reasonable, it is reasonable to act on them. Members can fulfill their obligations by reasonably believing that others won't do their share.

What this shows is that the collective can be blameworthy for failure, without it being true that the members are blameworthy for failure.

It might seem like there *must* be something strange going on here, for this to be true. If an obligation distributes, then surely when the distributed parts of it are done, the whole thing should be done. To give a simple example, if I have a cake, and cut it into eight pieces and give them to eight different people to eat, there shouldn't be any cake left. How can it be that each member of a collective satisfies her distributed obligation, and yet the group does not satisfy its? Strange as this may seem, it has to be the right answer. To get the distribution so that satisfying it *guarantees* the collective obligation being satisfied I'd have to make the shares categorical. I mentioned this possibility early on. But that would have the bizarre result that even when a member of the group *knows* that his contribution will be futile, and maybe even counterproductive given that he could be doing something else, he'll still be obliged to do it.

So for example, Kewa would be obliged to stay behind, pushing at the piano, even when Tom, Mark and Jonno had fled the scene. But surely he should be looking for others to help him, instead of doing something he knows to be futile. His doing his share in the absence of the others will not result in the collective action being produced. We only have the idea that the distributed obligations should be categorical because we want

the collective action to be produced. But making them categorical doesn't do that. Only people always doing what they are obliged to do would do that. Making the obligations conditional upon beliefs at least avoids the outcome that one person futilely or counterproductively does his share.

7.5 Conclusion

The feasibility tests presented in Chapters 4 & 5 seem to accommodate collectives as agents about which we can make 'feasible-for' assessments (and include within the scope of 'feasible-that' assessments), so long as we accept a particular story about what a group obligation means for each member, and when a group has an ability. Understanding the relation between group and member obligation lets us know *which* obligations or requirements need to be subjected to feasibility tests. Understanding group abilities allows us to rule actions in and out of collective agents' option sets, which is a precursor of running the comparative tests. The interesting thing to come out of looking at whether the feasibility tests accommodate collective agents is that blameworthiness can come apart between groups and their members. A group can be blameworthy when its individual members are not blameworthy at all.

Chapter 8

When is collective action likely to succeed?

8.1 Introduction

An outcome is feasible for a collective when some action that could realize it is in the collective's option set, and an outcome is more feasible for a collective than another when an action in the collective's option set is more likely to bring the outcome about than it (or some other action) is to bring another outcome about. Hard constraints rule actions out of option sets, soft constraints make actions less likely to produce desired outcomes. When we run agent-relative feasibility tests, we conditionalize on the agent's choosing (trying, doing) the action, and look at what soft constraints, including other people in the context, make likely to happen. This is a chapter about the 'other people' part of the feasibility story.

The bulk of the chapter must go to settling the terms on which we should talk about an outcome's likelihood of success. Once we have settled this matter, we can go on to trying to establish some general conditions that make success more likely (conditions which, if we want to make collective action more likely to succeed, we should try to make obtain). There are two main ways to talk about likelihood of success with respect to collective action. The first is in terms of what groups as groups are likely to do. We can predict an individual's actions when we know her preferences; the same should be true for groups. But that means to talk about the likelihood of successful group action we need to know about group preferences. And that in turn requires establishing whether groups are even the kinds of things that can have preferences (or can some kinds of groups have preferences, and others not?), and then talking about how group preferences are determined. People act in the pursuit of what they prefer; if groups are the same, and if we know what they prefer, then we can predict how they will act. And that opens up the possibility that we can start talking about how to manipulate their preferences so that they will act differently.

The second way to talk about the likelihood of success is in terms of individuals as group members. If that is possible, then we can skip over the discussion of whether groups can have preferences and if so how they are determined, and get straight into trying to identify the conditions that make individual group members likely to do their parts in producing group actions. Section 8.2 will be dedicated to the first possibility and

Section 8.3 to the second. I will argue that in the end, the two are equivalent: group preferences are *determined by* group members' preferences, and we can predict successful group action either by generalizing from the likelihood of individual contributions, or by talking about the likelihood of the group's acting in a way that realizes its goals directly. In Section 8.4 I want to discuss a possible exception, namely *groups in supergroup contexts*.

To reiterate, when we ask agent-relative feasibility questions, we ask about what an agent can do *in light of* others as constraints. This chapter is about others as constraints. When are the other members of a group likely to act? When are other groups likely to act? General answers to these questions supplement feasibility assessments by providing more content to context-relative soft constraints.

8.2 The likelihood of trying: groups as groups

Are groups the kinds of things that act intentionally? Only if they are can we talk about what they might be expected to do in certain kinds of situations. An irrational actor is an unpredictable actor. In Section 8.2.1 I shall introduce the idea of group intentionality via a discussion of states and corporations, two of the larger and more complicated kinds of groups, and in Section 8.2.2 I will ask whether we can just use Game Theory to predict successful collective action, given that groups are intentional agents. In Section 8.2.3 I will ask whether we can make sense of the notion of collective preferences, and whether and how we can derive collective preferences from individual preferences. In Section 8.2.4 I shall argue that there is a presumption in favour of talking in terms of collective preferences and collective action, due to the fact that group preferences are not reducible to individual preferences. In Section 8.2.5 I shall present an account of a tension between public and private preferences, as an analogy to the preferences an individual has alone, and her preferences in a group context. I shall argue that we can resolve this tension in a way that allows us to talk about group preferences, by talking about group-relative individual preferences.

8.2.1 From intentionality to rational action[42]

Alexander Wendt has asked about group agency for the specific group the state

42 Anywhere in this chapter that I talk about, or cite people talking about, 'utility', I mean this to be understood interchangeably with 'value' or 'desirability'. See also footnote 36, p. 146.

(and so his conclusions, if they are sound, do not extend to all groups but only to states or state-like groups). He discusses 'agency' by way of 'personhood', asking what qualities are in general sufficient to personhood, and whether the state can be said to have those properties. He focuses on three generally accepted conditions of personhood: having intentionality, being conscious, and being an organism. He distinguishes psychological persons from legal and moral persons, concentrating on the former. He begins with a rationalist model of psychological personhood. He argues that a state easily meets this criteria, and thus can be said to have intentionality. But he thinks a state fails to count as an organism, although can arguably be classified as a super-organism, and almost certainly fails to count as independently conscious.

> 'Rational actors' have four main properties: (1) a unitary identity that persists over time; (2) beliefs about their environment; (3) transitive desires that motivate them to move; and (4) the ability to make choices on a rational basis, usually defined as expected-utility maximization. These properties mean that persons are above all intentional – purposive or goal-directed – systems (Wendt, 2004, p. 295).

The question of how group intentionality is related to the intentionality of its constituent members can be answered by three competing theories: reductionism, supervenience, and emergentism. Reductionism argues that group intentions are nothing over and above the intentions of the group's individual members. This has classically been understood to be the view most compatible with physicalism, but it requires anti-realism or instrumentalism about groups as independent entities. Wendt comments that it would be something of a miracle if, given the usefulness of understanding states as persons, there weren't more to be had than biting the bullet on reductionism. Emergentism is the theory that some higher level properties are not reducible to or determined by lower level properties. But emergentism is too strong, because it maintains that what scholars in the area call 'I-intentions' are not obviously more fundamental than 'We-intentions', and that the content of thought (including collective thought) may be provided in part by context or environment.

Wendt discusses the example of a military unit in which no one person has full information about what each other person is doing, yet which is still able to fight a war. It looks in some cases like collective actions and intentions cannot be reduced to the actions and intentions of member parts. Between a reductionism that is too weak (because it settles for 'useful fiction' rather than serious ontological status), and an emergentism that is too strong (because it wants to make the group intentions fundamental), Wendt settles for the third option, namely supervenience. He takes this to

be a good way to go because it does not give into reductionism, yet it allows for the separate existence of the state as a group without making any weird metaphysical commitments. Supervenience works in one direction, namely from the bottom up. Any two states for which the micro-facts are the same (e.g. states constituted by the same members, who have the same intentions or mental states), will have the same macro-facts, namely group intentions, group character and so on. But that isn't true in the top-down direction because supervenience allows multiple-realizability. That means two states that are identical at the macro-level may have different supervenience bases, i.e. different micro-composition.

On Wendt's argument, states have intentionality, they may be organisms insofar as being a super-organism is close enough to being an organism, and they do not have consciousness. If all three conditions are necessary to personhood (agency), then the state is not a person (agent); but if one or more are sufficient, it is. Two things are interesting for our purposes, first, whether Wendt's discussion of states extends to other organized groups (e.g. he comments that 'corporate intentions ... are possessed by groups with a centralized authority structure capable of imposing binding decisions on their members' (Wendt, 2004, p. 297; see also French, 1979)),[43] and two, that intentionality was what we were interested in anyhow, not consciousness, organism-hood, or personhood in general. We don't need groups to be maximally like people in order to understand them or predict their actions and the potential obstacles to their actions. We can have 'theories of mind' about machines and animals after all; we just need to have some understanding of how they are programmed, or what is in their interests, or what their beliefs and desires are. So long as we have that, and we have observational evidence that they act to pursue those things in a consistent and deliberate manner, we will be able to say something interesting about their likely behaviour. The worry is that certain kinds of groups won't be rationally interpretable at all, in which case we won't be justified in expecting anything from them, nor in cooperating with them, if we can avoid it. It would be good if we could figure out how to tell that latter kind of group from

43 Peter French argues that corporations are intentional agents. They have explicit decision-making and ratification procedures. Those procedures subordinate and synthesize individuals' decisions in a company into one corporate decision. He sees organizational charts as the 'grammar' of corporate decision-making, and argues that we can understand corporate decision-making as a game played according to certain rules, with certain individual 'players' filling certain roles, which in turn stand in various relations to one another. French argues that if the corporate act is consistent with the corporate policy (the 'rules of the game'), then it can be seen as an intentional act. Corporations' reasons for acting are to further the interests of their long-term stated goals, regardless of the transient or conflicting goals of their personnel. Corporations, one kind of collective, have in that way beliefs and desires, or at least the functional equivalents of them (French, 1979).

groups with intentionality (more on this below).

8.2.2 Can't we just use Game Theory to predict successful collective action?

One way to try to approach the difficult problems of predicting successful collective outcomes is through mathematical models of rational decision-making. Rational Choice Theory, for example, takes the decision environment of a given individual as fixed, and on the basis of information about the individual's preferences can give us a good idea of what we can expect from her in terms of action. Game Theory does a similar job, but for individuals in contexts of strategic interaction. That is to say, it takes the decision environment as dynamic rather than fixed, allowing that the best action for a given player may depend on what the other players do, and what the other players do may depend on what they think the given player will do. Agent-relative feasibility assessments will often take place against this sort of background. Kewa's contribution to the piano being lifted will only be efficacious in the piano being lifted if the others contribute too; Iran's contribution to multilateral nuclear disarmament will only be efficacious in achieving multilateral nuclear disarmament if other countries disarm too. How can Kewa, or Iran, generate reasonable beliefs about what the others will do? Many of the difficult issues in theorizing about collective action, whether in predicting the likely success or failure of a group in coordinating their action, or predicting the likely success or failure of supergroup action (groups coordinating their action with other groups, each of which face in addition the internal coordination issues just mentioned), are a result of the fact that collective action involves the strategic environment modeled by Game Theory. So perhaps Game Theory can tell us something about the kinds of problems likely to face groups trying (or lacking a means) to coordinate their action, and the ways, assuming there are some, of overcoming those problems.

The traditional account of rational choice was that rational agents act in such a way as to maximize value. They have certain desires, and certain beliefs, and to act rationally is to maximize the chance of realizing their desires, according to their beliefs. For example, a person who desires an ice-cream on a hot day, and believes that there is an ice-cream shop four blocks away from where she is, would maximize the chance of satisfying her desire for ice-cream by walking the four blocks to the ice-cream shop and

purchasing an ice-cream. Simon Blackburn (1998) reverses this traditional story by arguing that rather than understanding the Principle of Maximizing Expected Utility as a rational requirement upon action (i.e. as normative), we should understand it as *definitional*: 'a grid imposed upon the process of interpreting others' (Blackburn, 1998, p. 161). We can interpret people as having an interest in an object when the object plays a role in their decision-making (Blackburn, 1998, p. 162). We can figure out what agents prefer *by seeing what they choose* (at least, in non-noisy conditions).

Of course, there may be a difference between an agent's preferences and the choices she actually makes in contexts where some things are outside her control. An agent may choose in a way that makes her 'safe' given her expectations about what others will do, and that choice may not be the one she would have made under more ideal conditions (Blackburn, 1998, p. 163). Preferences, then, are logical constructions out of choices given beliefs, and utilities (values) are logical constructions out of preferences. Any agent whose preferences are rationally interpretable can be interpreted in terms of utilities. It might seem at first glance that this interpretation of utility maximization as definitional rather than normative ruins any chance of using decision theory predictively. On the traditional account, we could take an agent's beliefs and desires as input, and use the principle of maximizing expected value to get an output, something that we expect the agent to do insofar as we expect her to act rationally. But on Blackburn's story, we can only say in retrospect that what she ended up doing *must have been* what she preferred, because on Blackburn's story, we assign preferences to people by seeing how they act. But it isn't entirely true that the account limits us to confirmation of rational choice only in retrospect. We are allowed to infer, on the basis of past choices and actions, what a person's preferences are. Thus we might come to have enough evidence about an agent's preferences, and about her various dispositions, to make accurate predictions about what she will choose in the future. We don't take her beliefs and desires as input, as on the traditional model, but rather we take her choices as input, and generate predictions using these in conjunction with certain reasonable inferences and assumptions.

None of this is immediately applicable to groups. Rational Choice Theory involves subjects making decisions in fixed environments, and Game Theory involves subjects making decisions in dynamic environments. But as the decision problems are standardly modeled, both are about individuals making decisions. Neither obviously model what a group should do, or can be expected to do, given certain facts about its

preferences. But as it turns out, this doesn't matter at all. Game Theorists produce mathematical models, and so long as a player is rationally interpretable, it doesn't matter what kind of thing the player is, human or otherwise. Game Theory is used to model the strategic decision problems of individuals, but it is also used to model the decision problems of countries, animals, corporations, and so on (Ross, [1996] 2006, p. 6). So long as I can establish that a given group is rationally interpretable, there should be no problem in using Game Theory to make predictions about group behaviour.

Theorists in the game theory literature have been rather bold in saying what we should 'expect to see' from players in certain game contexts. For example:

> [In the fairness experiments, s]ince the game is played only once and the players do not know each other's identity, a self-regarding responder will accept any positive amount of money. Knowing this, a self-regarding proposer will offer the minimum possible amount, $1, and this will be accepted (Ross [1996] 2006, p. 177).

...and:

> [In the Prisoner's Dilemma, t]herefore, you're better off confessing regardless of what she does.... In the PD ... confessing strictly dominates refusing for both players. Both players know this about each other ... Thus both players will confess, and both will go to prison for five years (Bowles & Gintis, 2006, p. 11).

The problem with these kinds of predictions is that they do not make explicit that certain assumptions about players' utilities are being made. As already explained, players, so long as they are rational, can be interpreted as having acted on the basis of their preferences. To predict what they will do in advance, we have to know what their preferences are. The two predictions above rest on the assumption that human individuals are exclusively self-interested. *If* a player in a fairness game is exclusively self-interested, *and* knows the game is both one-shot and anonymous, *then* she can be expected to offer the minimum possible amount to her co-player. And *if* a player in a Prisoner's Dilemma is exclusively self-interested, *then* he can be expected to confess.

But many people are not exclusively self-interested. People prefer many and varied things. As Don Ross puts it, 'a utility function for a player is supposed to represent *everything that player cares about*, which may be anything at all' (Ross, [1996] 2006, p. 19). Or according to Samuel Bowles and Herbert Gintis:

> Preferences are reasons for goal-oriented behaviour. Preferences thus include a heterogeneous melange: tastes (food likes and dislikes, for example), habits, emotions (such as shame or anger) and other visceral reactions (such as fear), the manner in which individuals construe situations (or more narrowly, the way they frame a decision), commitments (like promises), socially enforced norms, psychological propensities (for aggression, extroversion, and the like), and one's affective

> relationships with others. To say that a person acts on her preferences means only that knowledge of the preferences would be helpful in providing a convincing account of the actions – though not necessarily the account which would be given by the actor, for as is well known individuals are sometimes unable or unwilling to provide such an account (Bowles & Gintis, 2006, p. 174).

If a player in a fairness game cares about fairness, then we should expect him to propose a more egalitarian distribution of the endowment, just as if a player in a prisoner's dilemma -type situation cares about the welfare of the other prisoner, we should expect him to remain silent. I say 'prisoner's dilemma -type' instead of 'Prisoner's Dilemma' because it is important to notice that some games are defined by the utilities assigned to the players. That is to say, a player is only *in a Prisoner's Dilemma* when he faces the payoff matrix given in that game, namely, doing better according to his own preferences by confessing, no matter what the other prisoner does. Ross reiterates this point:

> In general, then, a game is partly *defined* by the payoffs assigned to the players. If a proposed solution involves tacitly changing these payoffs, then this 'solution' is in fact a disguised way of changing the subject (Ross, [1996] 2006, p. 19).

This is an important point, because many writers have carelessly supposed various empirical experiments to have refuted the predictions of game theoretic models. But these models only make predictions on the basis of certain assumptions about players' preferences. Divergence from the predictions of the model only shows that the wrong model was used.

The upshot of all this is that Game Theory won't give us anything for free in terms of predictions about group action. If we know that a group is exclusively self-interested we might exploit some of the mathematical modeling that has been done assuming this account of an agent's preferences; if we know that a group has preferences for the welfare of others roughly along the lines exhibited in fairness experiments we might exploit some of the empirical research that has been done there. But the models of preferences that economists have engaged with are limited, and the experimental data only tests a small number of games (the biggest literature exists for ultimatum and public goods games). Thus what we really need to know if we want to assess the chances of successful group action is what the group *prefers*. But this just opens the door to a host of new questions. Is there any such thing as a group preference? Are groups rationally interpretable in the same way that individuals are, so that it makes sense to talk about what they prefer? If so, what is the relation between group preferences and individual preferences? What kinds of problems can confront groups to either distort their

preferences, or make it the case that their preferences do not translate into action in the straightforward way they do with individuals?

8.2.3 Group preferences and rational interpretability

In this section, I want to ask whether there is any such thing as a group preference, canvassing several accounts of group preferences in the growing literature on group action. This will involve discussing the relation between individual and group preferences, and the generally accepted claim that group preferences are not reducible to the preferences of individual members, even though group preferences supervene on individual preferences. Through this discussion, some of the unique problems confronting collective action should become more clear, although of course any detailed treatment will require focusing on a particular kind of group. Having shown that the notion of a group preference is a coherent one, I will sketch the requirements of rational interpretability. A group's action (and dispositions, as we interpret them) must meet certain requirements if we are to interpret their behaviour as rational or goal-directed; this is what distinguishes the group acting purposively from e.g. a tree falling over in a storm. I will show that some groups fail miserably in meeting these requirements of rational interpretability, which indicates that we cannot assume merely from the fact that a group exists that we can interpret their behaviour in any rational way. Perhaps surprisingly, this is especially true of democratic states.

In his (2000), Robert Sudgen defends the idea that groups can have preferences that are distinct from the preferences of each of the group's members, even to the point that the group preference is something that *no* member of the group prefers. He gives an example of taking holidays with his family. As a family, the Sudgens prefer trips where there is scenery and wildlife. This is true of the family even if for example Sudgen individually prefers trips to big cities, his wife prefers trips to remote spa locations, his daughter prefers trips to wherever the shopping is good, and his son prefers trips to wherever there are likely to be attractive girls. Furthermore, the preferences of the family function in roughly the way preferences function in the decision-making of individuals:[44]

44 There are two separate issues. One is how the group preference is constructed out of the members' preferences. The other is whether the group acts on its preference. The first issue is more difficult. With the second issue, the group preference and the individual members' preferences will not come apart so long as each member prefers to do what the group prefers to do.

> ...the combined effect of the choices of the members of the team will be to bring about the outcome which, of those that are feasible, is most highly ranked in terms of the team's preferences. So it is as if the team were a single agent, choosing among feasible outcomes according to its preferences. In this sense, it is meaningful to talk about the team as an agent in its own right (Sudgen, 2000, pp. 196-197).

He argues that it is important that members understand themselves to be members of a particular group, engage in team-directed reasoning, and have confidence that the others in the group will be engaged in team-directed reasoning. For example, a football team is composed of players who understand themselves to be members of that team, who reason about action in a team-directed way, e.g. 'our objective is to score as many goals as possible', and who have confidence that their fellow players are engaged in similar reasoning. That is why a football team can have the goal of scoring as many goals as possible, without that necessarily being the preference (although it normally will be) of a particular player.

Margaret Gilbert argues along similar lines that group preferences come from goals or expressions of willingness to be bound in a shared commitment. A person expresses their willingness to others in conditions of common knowledge, and then is only released in negotiation with the group, which is to say, a sole member cannot change the group's mind about its goal (Gilbert, 2001). Gilbert thinks that group preferences create individual obligations in this way; once a person has expressed her willingness to be bound in a shared commitment, and others have accepted in conditions of common knowledge, she is *bound* to that shared commitment. On this point, Gilbert and Sudgen disagree; Sudgen denies that the mere existence of collective preferences is sufficient to the generation of individual obligations.

Relatedly, Gilbert thinks that a group can *believe* something without a majority of group members believing it, or even any of the group believing it. She proposes a joint acceptance model, on which something becomes the group belief because others accept it, or let it stand as the group belief when it is put forward, rather than objecting to it. Preferences might work in much the same way, so that the collective preference might be determined by a suggestion made by a member of the group that no other member assents to (perhaps because those members believe that all the others assent) (Gilbert, 1987). Gilbert thinks that states are a complicated case for the joint acceptance model, but that the model can be extended to fit them. She suggests that the government's view can represent our own because their view is tacitly accepted to reflect our own, but in a way that is defeasible (Gilbert, 1987, p. 200).

It is worth mentioning that there might be some ambiguity in our language when we talk about collective beliefs and collective preferences. If my friends from New Zealand visit me in Berlin, I might ask myself 'what do they prefer to do this evening?' or 'do they *really* believe that the best way to experience Berlin is to stay indoors eating borscht?' And I might mean to ask what they, as a group, prefer to do, and believe. But I might also mean to ask a question about *the majority* of the group, or almost all of the group, asking what they each prefer, and believe, and how that adds up into some sort of group preference and group belief. It seems plausible that I might say 'my friends want to go out to a bar tonight', even though only one of them is enthusiastic and the others indifferent. It also seems plausible that I might say 'my friends aren't really keen to go out tonight', because three of them would rather stay in, even though one of them would really like to go out. In this chapter I am concerned with the sense in which the preference is really *the group's*.

In some cases, the goals of a team are a function of the kind of team they are; it would hardly be a football team whose objective was to score the *least* amount of goals possible. This may be true in general of sports teams, companies, clubs and societies, etc. But it clearly isn't true of lovers or friendship groups, on the other hand. Neither Sudgen nor Gilbert say *how* individual preferences are transformed into team preferences. What is the relationship between the one, and the other? I can imagine several answers to this question; perhaps individuals rank their preferences in a numbered table and simply pick the option they all assigned to the same number in the table. For example, if all members of the Sudgen family rank scenery and wildlife as fourth out of a possible ten options, and there is no unanimous agreement on any of their first three ranked options, then we might say the Sudgen family prefers holidaying where there is scenery and wildlife. But in that case, there could be an option which all family members rank higher than fourth which is nonetheless not chosen because of a failure of consensus. Or worse, the bizarre outcome that a group who could agree on nothing but which option should be ranked lowest, by virtue of that sole consensus find their group preference to be that which they all agree they least prefer. If we use majority rather than consensus voting, we get something like the electoral voting system in Australia, where a person's vote is transferred to candidates lower and lower in her list as those candidates are eliminated from the race, so that in theory, if everyone ranks a candidate in their top three, but differ on the position they assign to her, that candidate may win the election despite not having a majority of first position votes.

Of course the most obvious way in which we might move from individual preferences to group preferences is by simple aggregation. Assume that every member of the group gets one vote as to what the group prefers, and the group preference is determined either by a simply majority (50% or above) or by a supermajority (arbitrarily higher, e.g. 75%). That is just to appeal to the majority sense of collective preference mentioned in the last section. What's wrong with this way of determining collective preferences?

The short answer, made famous by Kenneth Arrow in his work on voting systems, is that individual preferences cannot be aggregated into a group preference without violating important constraints. These constraints are (1) non-dictatorship, that one person cannot decide the preferences of the group on behalf of the group (this is sometimes discussed under the heading of 'symmetricity', that every person's vote counts for the same amount); (2) the Pareto condition, that if there is an outcome *Y* that everyone in the group prefers to outcome *X*, the group's preference order should rank *Y* above *X*; (3) irrelevance of independent alternatives, that if *B* is preferred to *C*, introducing *A* should not make it turn out that *C* is then preferred to *B*; (4) transitivity, that if the group prefers *Y* to *X*, and *W* to *Y*, then the group prefers *W* to *X*; and (5) universality, that every set of individual preferences should result in a complete group preference, and the same set should produce the same group preference on any occasion. In the case that there are only two alternatives to choose from, majority rule succeeds in creating group preferences out of individual preferences that satisfy the above constraints, but whenever there are more than two options, the rule fails (see Arrow, 1950; 1951; 1963). For any decision procedure designed to aggregate individual judgements into a collective judgement, when there are more than two alternatives, it will have to treat some individuals, or some issues, as more important than others (in the worst case requiring a dictator), or will have to let the collective view on an issue be determined by the collective view on other issues (see e.g. List & Pettit, 2002; List & Pettit, 2004). Brian Weatherson uses Arrow's constraints as conditions of 'rational interpretability', which is to say, an agent is rationally interpretable only when his preference order conforms to those constraints (Weatherson, ms. p. 118).

Thus the straightforward aggregation of individual judgements into a collective judgement using a majoritarian decision procedure is highly problematic (it seems to me that groups with dictators may well be rationally interpretable, the problem here is with groups that want to respect Arrow's constraints *and* use majoritarian decision

procedures). We cannot rely on groups using majoritarian decision procedures to be rationally interpretable, and thus predictable. If we want to understand a group as rational (or if the group itself wants to *be* rational) we must derive the group preference in some other way than by aggregation of individual preferences (or the group must find another way to determine its own preferences). I have established that groups are indeed the kinds of things that can *have* preferences, even though we have seen that there is some difficulty in figuring out how to move from individual to group preferences. Perhaps the constraints above are too strong. Let's see what other writers suggest.

Decision theorists usually distinguish decisions under uncertainty from decisions under risk. A decision is taken under uncertainty when the agent has no beliefs at all about what the relevant probabilities are. A decision is taken under risk when she does. Decision theorists have argued a lot about the various decision rules which should govern the choosing of some action over another under conditions of uncertainty. These include maximin (maximizing minimums, i.e. choosing the option with the best worst outcome), minimax regret (minimizing maximal regret, i.e. choosing the option that would produce the least regret), maximax (maximizing maximums, i.e. choosing the option with the best best outcome), optimism-pessimism (weighting maximin against maximax), and insufficient reason (assume that each state is equally probable, and choose the act with the greatest expected value). Most of these rules require that a person's preferences be invariant under positive linear transformation (when the intensity of a preference ordering is expressed using an interval scale, what matters is not the numbers assigned to each but the fact that they respect the intervals, so many linear transformations will be possible). That is to say, for most of these rules, the preferences must be *cardinal* rather then merely *ordinal*. The maximin rule is the exception, it works even when preferences are only ordinal (Resnik, 1987, Ch. 2).

However, decision theorists' distinction of 'decisions under uncertainty' from 'decisions under risk' is a red herring. We assign utilities according to *subjective beliefs* about the probabilities of outcomes occurring. Very rarely, if ever, do we have no beliefs at all about a situation. Of course we sometimes have imprecise beliefs about probabilities due to a lack of information, but in that case we will generally be able to specify a range of probabilities (for example, between 0.3 and 0.5), which is still more information than none at all, or we will hover around 0.5. And furthermore, as we have seen already, none of the decision rules for these alleged situations of uncertainty are particularly good. The maximin rule is the favourite because apparently the least

demanding, but it asks that persons think of the worst *possible* outcomes for a set of potential actions, and choose the action with the least bad worst outcome. But think about what that would mean. It's possible that an event with extremely low chance will occur and have terrible consequences. If those consequences are worse for one action than for another, then we should choose the other (because we should choose the option with the least bad worst consequence), even if the other consequences, much more likely to happen, are worse overall for that action. Surely what matters is not how bad the consequences of the worst outcome are, but rather how likely each of the outcomes are. If an event is vanishingly unlikely, it shouldn't enter into our practical deliberations at all (but this comes back to the point about decisions under uncertainty, which assume absolute uncertainty). Thus in what follows I will borrow from discussions only about decisions under risk, which basically capture everything that's needed.

So what does a decision under risk need to look like for a person's choice to be rationally interpretable? To be rationally interpretable, preference orderings must satisfy certain conditions. These will differ depending on the exact version of decision theory (especially causal versions) being used, but there is substantial overlap (for a useful survey see Fishburn, 1994). The following are best known, given by Leonard Savage ([1954] 1972). They separate into two categories, 'structure axioms', and 'axioms of pure rationality'. Because some are extremely complicated, I will set the structure axioms aside in favour of his pure rationality axioms, and give only a very rough idea of the latter.[45] There are five axioms of pure rationality. These are *partial ordering, completeness, independence, nullity, and stochastic dominance* (Savage, [1954] 1972; Joyce, 1999, Ch. 3). Partial ordering includes *reflexivity of weak preference* (*x* is weakly preferred to *x* (where 'weak preference' means 'preferred at least at much'); and *transitivity* (if a person prefers *a* to *b*, and *b* to *c*, then she prefers *a* to *c*, and if she prefers *a* to *b* but is indifferent between *b* and *c*, then she prefers *a* to *c*). Completeness requires that between any two options, decision makers weakly prefer one over the other. Both independence (a preference between *a* and *b* should not depend on the circumstances in which the two produce the same outcome) and nullity (if we are sure an event will not come about, then the fact that an agent would prefer some outcome to another *if* that event did come about should make no difference to her preferences) are parts of what is popularly known as *the sure thing principle.* Stochastic dominance requires that an agent prefer prospects that

45 I follow the presentation of Savage's conditions in Joyce (1999, pp. 97-113). Joyce tends to present the conditions in terms of mental states rather than in the behaviourist terms favoured in revealed preference theory.

offer her a greater chance of obtaining the more desirable outcome (Savage, [1954] 1972; Joyce, 1999, Ch. 3). Three of these show up in all accounts: transitivity, reflexivity and completeness (see e.g. Blackburn, 1998, Ch. 6; Dreier, 2003, p. 160; Resnik, 1987, Ch. 2; Ramsey, 1931). The five conditions are sufficient to rational interpretability, but not necessary; several can be violated while an agent is still rationally interpretable. We cannot do without those that are not necessary, however, because we would then lose sufficiency. What is needed is a weakening of the non-necessary conditions, but it has proved extremely difficult to find one.

Whenever a set of group preferences does not satisfy partial ordering, completeness, independence and nullity (the sure thing principle), and stochastic dominance, then we cannot make predictions about what the group will do, because the group is not rationally interpretable. But conforming preferences to these axioms is not overly demanding, so there is no reason to think groups should not be rationally interpretable, at least in some cases, some of the time.

8.2.4 Presumption in favour of establishing likelihood using groups as groups

In the last section we considered Sudgen's and Gilbert's claims that a group can have a preference that *no member of the group has*, or that *the majority of the group doesn't have*. This is possible because their models of group preferences are based on assent, or team-directed reasoning. Gilbert's example was that a member of the group might assert a preference for the group, believing that it reflects what the others prefer even though not what she herself prefers, and the others may each assent to it (even if only by failing to object) for the same reason. Sudgen's example was that his family's preference with respect to its holiday plans is different from each family member's preference. What these examples seem to show is that group preferences are *not reducible to* individual preferences. If we wanted to build the Sudgen family's collective preferences out of the preferences of the individual family members, we'd have to somehow get from the conjunction of individual preferences (big city & remote spa location & good shopping & attractive women) to the family preference (scenery and wildlife). It's hard to see how the latter could fall out of the former. Likewise, we couldn't build a group preference in a case like Gilbert's out of a conjunction of individual preferences *none of which* are the group preference (because then even the 'dictator' rule, in which the preferences of the

group are decided by one member on the group's behalf, cannot make sense of the end result).

This tells us there should be a strong presumption in favour of talking about the likelihood of successful group action in terms of group preferences rather than in terms of individual preferences. Group preferences are distinct from, and not reducible to, individual preferences, which means that if we try to predict group action on the basis of individual preferences, we will get distorted results.

Can that be right? In the next section I want to introduce a distinction between public and private preferences, by way of analogy to the non-reducibility I have just claimed. The objection to that distinction should make clear the objection to this claim, and the idea that there should be any presumption in favour of dealing directly in collective preferences. I hope to show that it makes little difference whether we deal with the preferences of individuals (as group members), or directly with collective preferences – the predictive upshot should be the same.

8.2.5 Public and private preferences? An analogy

In a series of works beginning at the end of the 1980's, Timur Kuran presents a model of preferences in which there is a tension between a person's 'true' private preferences, and her 'false' public preferences (see also Kinder & Kiewiet, 1981; Kiewiet, 1983; and Rohrschneider, 1988). Kuran argues that individuals derive utility from three sources: available options, social sanctions, and decisional autonomy. From these conflicting sources come two very distinct kinds of preferences; a person's private preferences, which are known only to her, and a person's public preferences, which she presents to others (Kuran, 1990, p. 2). Kuran argues that to be accepted and respected in a given society, a person must be seen to support its basic institutions, and to support its main objectives. He surveys a number of empirical experiments which show how people's beliefs are dependent upon their perception of others' beliefs, and their behaviour sensitive to others' behaviour, and argues that the positive and negative sanctions associated with a person's public preferences create 'reputational utility'. A person who prefers that the laws allowing gay marriage be repealed, but who wants to be accepted by a community of gays and atheists, might choose to keep her true preferences private (Kuran, 1990, p. 7).

But so far that only means that we have preferences that we sometimes do not

reveal, not yet that we have public and private preferences which come into conflict with one another. Where does the split come from? Kuran argues that the social pressures that create reputational utility trade off against decisional autonomy. Some people place a high value upon choosing for themselves, no matter how strong social pressures are. These people may even take social pressure to be a *reason* to decide otherwise. Full decisional autonomy comes from supporting in public the option one prefers most in private. Decisional autonomy is compromised by supporting an option in public that one considers suboptimal in private, and it is the latter that Kuran calls 'preference falsification' and is most concerned with (Kuran, 1990, p. 11). He thinks the total utility of a public preference (falsified or not) comes from adding the preference's impact on the community to the utility of the generated sanctions, and adding both of these to the utility derived from the extent of decisional autonomy exercised.

Individuals are tempted to falsify their preferences because they stand to gain in terms of reputational utility (even though the cost is decisional utility). His complaint against neo-classical economics is that it equates public with private preferences, when the two are genuinely distinct. Human persons are complicated; they gain reputational utility from their social standing, but personal utility from autonomous decision-making. Sometimes what they gain from falsifying their preferences will be sufficient to make up for the loss of decisional autonomy such falsification costs. The main departure from standard economic theory is in taking preferences not as fixed but as changing in response to social norms and standards.

Cass Sunstein, in a review of one of Kuran's works in which the view was defended, comments that the theory goes a long way in illuminating the prospects for social stability and social change: '[...a]s people's thoughts about other people's thoughts change, there is a shift in reputational incentives, and hence people's public preferences can shift: if you come to believe that there is a "silent majority" believing what you believe, you probably won't be silent for very long' (Sunstein, 1995, p.2). But he objects to the idea that people's private preferences are the true ones and their public preferences are the false ones, as Kuran suggests that they are. People may have racist, or sexist, or otherwise objectionable private beliefs and preferences, in which case their being malleable under social pressure is a good thing. In that kind of situation it's not clear that we should say their preferences have been 'falsified' by public pressure, or that their public preferences are not the 'true' ones.

This kind of objection can be taken even further (and this objection is how we

should also respond to the presumption in favour of predicting collective action only by way of collective preferences). Why think that a person's preferring differently in public than he does in private *means* he has two different and conflicting preferences, one of which he chooses to reveal, and one not? We can instead account for both the public and the private contexts within the one set of individual preferences. Preferences are distributions of utility over worlds. If they were simple, something like Kuran's story might be right. On his story, it might be true that in private Peter prefers pineapple to mango, and in public prefers mango to pineapple. But if Peter prefers *x* to *y* and also *y* to *x*, he fails rationally. Thus (if he is rational), only one of these must be his 'true' preference, and the other must be a falsification. On Kuran's story it is a falsification intended to raise the person's social utility. Peter is in a community of mango-lovers, so he does better by pretending to prefer mango. But why think that preferences are so simple, especially when that is the result? Preferences might rather be complicated disjunctions. Maybe Peter equally prefers being in a world where he is alone and eating pineapple, and being in world where he is with friends (let's say friends who are mango-lovers) and eating mango. Given that he is in a world with such friends, his preference will be for mango. That does not make him irrational or conflicted.

The point is that preferences are context sensitive, and they are seldom fixed. They change according to who we are with, what we think other people want, and so on. That was part of Kuran's insight, but he took from it a different kind of conclusion. We are not, by virtue of our complicated disjunctive preferences, thereby irrational or dishonest. That is just how things work for most people. (Notice that while you might not be sure when alone if you like a particular item of clothing, or piece of art, or song, or meal, you may love it when a friend whose opinion you respect declares that she loves it). People often *seem to* radically change their preferences when they get a new partner, or enter a new group of friends, or workplace. But rather than seeing that as a disingenuous falsification of the person's 'true' preferences, we should see it as just the way preferences function. One might not like football but still watch it a lot with one's partner (let's say because one prefers spending time with one's partner and watching football, to spending time alone and watching *Breaking Bad*), and one might find that even after that relationship is over, one watches football alone and genuinely enjoys it. Preferences are the kinds of things that evolve over time and context.

The same objection to Kuran's distinction between public and private preferences goes for the claim that group preferences are not reducible to individual

preferences. We should not simply think of individuals as having fixed preferences, which are somehow distorted in the group context. Rather, we should think of them as having preferences for things they might prefer to do alone, and preferences for things they might prefer to do in the group context, where they are sensitive to what others prefer to do. This can again be expressed disjunctively. Returning to the case of the Sudgen family holiday, we might say that Sudgen's wife prefers either the world in which she holidays alone and goes to a remote spa location, *or* the world in which she holidays with an old friend and they stay at a hotel in a small village, *or* a family holiday in which they go somewhere with scenery and wildlife. She might be indifferent between these three worlds, but the family holiday might be the most feasible for her at the time. Or she might prefer the family holiday over the other two. In that case, we should say that she prefers a *family holiday with scenery and wildlife*. That is not to say that she alone prefers a holiday with scenery and wildlife, it is to say that she prefers the conjunction of being with her family and having a holiday with scenery and wildlife.

None of this determines how a given group reaches a decision about what their preference is (although presumably if you know the other group members well enough it's easy to figure out activities and actions that all will be happy enough with, given that they prefer to be in the group and do something over not being in the group). But it does say that individual preferences are not totally different things to group preferences. Individuals have *group-oriented* preferences, preferences that are sensitive to their being in a *group context*. If we know enough about the group, then we can build group preferences out of individual preferences after all.

8.3 The likelihood of trying: groups *qua* individual members

Agent-relative feasibility assessments are sensitive to constraints of context. Whether I can contribute to producing a successful collective outcome depends on what the other members of the collective are likely to do. But what are they likely to do? Is the likely success of a group action *determined* by the likelihood of group members choosing to do their parts? What does it take, exactly, to determine that a group is likely to succeed in its action? What are the particular considerations about its members that we should entertain?

8.3.1 From individual preferences to group preferences

In the final subsection of the last section, I argued that Kuran's distinction between public and private preferences rests on too simplified an understanding of preferences. If Kuran allowed for conditional and contextual preferences, the dichotomy between public and private preferences, not to mention the implicit assumption that private preferences are the 'true' or 'real' preferences, would collapse. I argued that the same objection we make to him can also be made against the claim that group preferences are not reducible to individual preferences. On a very simple understanding of preferences, it can be that a group prefers something different to what its members each prefer. But on a more detailed account of preferences, the two can be reconciled.

In Sudgen's case, he prefers the possible world in which the family take a holiday together *and* go somewhere with scenery and wildlife, to the possible world in which he takes a holiday alone *and* goes to a big city; and so on *mutatis mutandis* for the other members of the family and their respective individual preferences. People have preferences with respect to what they like to do alone, but they also have preferences with respect to what they like to do with others. Perhaps Sudgen would rank the world where the whole family goes on holiday to a big city *and enjoys it* the highest, but knowing as he does the individual preferences of the rest of the family, and in particular his wife, he perceives that his choice is rather between the family going on holiday to a big city *and hating it*, the family going on holiday somewhere with scenery and wildlife *and enjoying it*, and his going on holiday to a big city alone. So long as he prefers the family holiday in the compromised location to his holidaying alone, and so long as he see that some of his preferences are unrealistic (such as the preference that the family go to his preferred location and enjoy it), we can take the family preference to be his group-oriented (or contextually sensitive) preference. Thus there is no problem in saying that group preferences are reducible to individual preferences, we just have to be more careful about how we think individual preferences work.

8.3.2 An equivalence: groups as groups, individuals as group members

If group preferences are reducible to individual preferences on a more detailed understanding of individual preferences, then it doesn't matter which way we come at

the issue of predicting successful collective action. If group preferences couldn't be reduced, we would have been forced to talk directly about group preferences, which would have required figuring out how we can come to know them (although observational evidence is one source; if we take Blackburn's or Dreier's analytic accounts of value maximization, we can infer what people prefer from what they do, so we could infer group preferences from group action), and then making claims about what should follow from them. But they can be reduced. That doesn't mean we should only ever work upwards from individual preferences, but it does mean that nothing much should hang on how we choose to talk about the issue.

We can talk in terms of group-oriented individual preferences, or we can talk directly in terms of group preferences. The question that remains is, what *are* those preferences, and what do they tell us about the chances of successful group action? In the next section, I want to concentrate on individual contributions (if we don't know what a group prefers, we can either infer it from behaviour as already mentioned, or we can build it out of individuals' group-oriented preferences). People generally do what they prefer. If all the members of a group are such that they prefer to contribute to the group action, then they will generally do so, and thus the group action will generally succeed. So what are the conditions that make an individual less likely to contribute, because more likely to prefer not to? And relatedly, can we manipulate any of those conditions to make cooperation more likely?

8.3.3 Under what conditions is individual contribution likely?

I have used Sudgen's example of a football team to illustrate the problems of talking about group preferences. But sports teams might just be an easy kind of case, given that they have obvious objectives that are part of what kind of group they are. What do we do about the more difficult kinds of groups, that don't have any fixed objectives? Where do we even start, in figuring out what they prefer, and thus when action is likely? It seems that in such a context we have only a few options. We can *ask them* what they prefer and hope for an honest answer, we can watch them (observational evidence and inference by induction to future occasions), or we can just make generalizations from what people *in general prefer* (or more manageably, when people in general will fail to act). We can't do the latter without having to make certain assumptions about human psychology and behaviour, but if these assumptions are true

in a majority of cases that should not be objectionable.

In this section, I want to make a start on the last option, suggesting some conditions under which an individual in her role as a group member will be more likely to contribute to producing a collective outcome. Of course these conditions will be general (there's no accounting for some people's preferences), but they should be plausible as defeasible conditions upon the likelihood of success of a group action, determined by the aggregate likelihood of success of the group members' contributions to the group action.[46] Let me simply list them here, and then go on to discuss each in more detail.

(1) Privilege.
(2) Perceived Categorical Moral Obligation.
(3) Preference for Success.
(4) Joint Necessity.
(5) Salience.
(6) Non-harm.
(7) Tipping point: non-contribution.
(8) Tipping point: contribution.
(9) Difficulty.
(10) Informedness.

(1) *Privilege.* In his seminal work *The Logic of Collective Action*, Mancur Olson argued for a coarse typology of groups into 'privileged' and 'latent' (Olson, [1965] 1971). He thought that these correlated roughly with 'small' and 'large' respectively, although I shall argue that this is not the case. A group is privileged if a person gets more back from a group activity than they invested in the activity, and latent if a person gets less back (or breaks even). For example, suppose that three individuals each work as contractors making coats for a clothing company. For a typical 40-hour working week, each of them produce on average six coats for the company. Now suppose they discover that each of them are skilled in different steps of the garment-making process. The first individual is a gifted designer; the second individual is a skilled pattern-maker and fabric cutter; the third individual is a talented seamstress. By dividing their normal labour according to

46 At this point by 'contribution' I mean their genuinely doing parts of the action, not simply satisfying the conditional obligation (to do a part, conditional upon the relevant belief).

specialization, suppose that one of two scenarios results.

In the first scenario, the group produces a total of twenty-four coats in their first week of work. In the second scenario, the group produces a total of fifteen coats in their first week of work. According to Olson's typology, the group in the first scenario is privileged, because each individual's contribution to the group activity, i.e. their standard week of labour, is increased by the group. Instead of the six coats produced when working alone, each individual (assuming equal division of goods) has now produced eight coats, which they can sell to the company at greater profit. The group in the second scenario is latent, because each individual's contribution to the group activity, i.e. their standard week of labour, is decreased by the group. Instead of six coats, they each produce only five by working together. Later empirical research has confirmed Olson's claim. In an exhaustive meta-study of approximately thirty years of public goods experiments, John Ledyard concludes that one of the two factors to have the greatest impact on making cooperation likely is increasing the marginal per capita return, i.e. the payoff in proportion to the contribution (Ledyard, 1995).

This distinction between types of groups explains in a straightforward way why rational individuals have an interest in sustaining privileged groups, but it doesn't yet explain why there should be any correlation between privileged groups and small groups, or latent groups and large groups, as Olson thought there was. His idea was that individuals can conditionalize their actions upon what others in the group will do much more easily in small groups than in big groups, simply because in small groups they will have greater knowledge of others' actions. Furthermore, small groups, he thought, are better able to utilize mechanisms like solidarity, moral suasion and strategic interaction to overcome latency, mechanisms less likely to be available to large groups. Thus, Olson argued, small groups are more likely to succeed in their enterprises, and large groups are more likely to fail. If this were true, it might be utilized politically, for example in arguing against the cosmopolitans in favour of the persistence of sovereign nations rather than a world government, and for nations of federated states, states of federated districts, and so on (see discussion in Hardin, 1982, Ch. 3).

Russell Hardin argues that the correlation Olson assumes between latent groups and large groups is unjustified. Just as there are cases where large groups are more costly to organize and more difficult to control (and therefore less likely to succeed in their enterprises), there are cases that go the other way, for instance where some or other of the group's goals is costly, which would be burdensome if divided among the members

of a small group, but manageable when divided among the members of a large group. So in some cases large groups are more likely to succeed, because there are more people to share the costs of obtaining a desired outcome. The important distinction, then, is between latent and privileged, not between small and large (Hardin, 1982, p. 38-43). It should also be noted that many of the mechanisms available to small groups thought to add to the likelihood of successful group action may be available to larger groups as technologies improve. Solidarity and moral suasion might be created through media like the internet, radio, television, social networking sites, advertising and so on, and strategic interaction might be virtual instead of physical.[47]

(2) *Perceived Categorical Moral Obligation.* When individuals perceive themselves to have a categorical moral obligation to do their part in a collective action (whether as a religious or cultural artifact, or as the result of targeted persuasion e.g. campaigning, fact-giving, reason), they are more likely to do it. This is not to say that people are likely to do what they are morally obliged to do, but that people are more likely to do what they *take themselves* to be morally obliged to do. Ralph Wedgewood defends the idea that it is a condition of rationality that, having judged that I ought to φ, I form an intention to φ (Wedgewood, 2007). This ties into the discussion of goods that involve incremental value rather than a threshold, in Chapter 7. If the people of Berlin take themselves to be obliged to contribute to Berlin's effort to donate $5 million dollars to aid efforts for the Pakistan floods, then the fact that they don't think *everyone* will make a contribution, and that therefore Berlin won't succeed in that effort, is not a reason for them to fail to contribute. Every contribution helps. The fact that it is an incremental good at issue rather than a threshold good means that a person is more likely to contribute (assuming she knows this) no matter her beliefs about others' probable actions.

(3) *Preference for Success.* This condition does not alone make contribution likely, but it works in combination with others, e.g. (7) and (8) below, to do so. This condition notices that people will be more likely to contribute, especially when other conditions are

47 Even if we adopt the rough idea that privileged groups are more likely to succeed in their enterprises than latent groups, there are some difficulties in saying exactly what constitutes a privileged group. In the stylized cases used to describe the differences between the groups, goods are fungible, which is to say, it is obvious when a person gets a return higher than their contribution. In the example we considered, the goods were the same (e.g. coats), and to further simplify matters, the case was comparative between working alone and working in a group. But what about cases where goods can only be obtained by working in a group? And what about cases where an individual's contribution is made in different units to the return (say she contributes time and energy, and gets out goods; or she contributes money, and gets out services)? Or cases where the output of the group action is the *avoidance* of some catastrophe or other, so we must assess her contribution against a return that is in some sense only valuable relative to a counterfactual?

met, when they *prefer* the conjunction of contributing and the group action succeeding, to the conjunction of not contributing and the group action failing. That is to say, they find the benefit of the collective good being produced to be worth the cost of their contribution.

(4) *Joint necessity.* Members of a group are more likely to contribute to some group action when they perceive themselves to be jointly necessary, along with the other members of the group, in producing the group action, and when they prefer group success as per (3) above.

(5) *Salience*. This is an epistemic condition. People are only likely to contribute to projects when the fact that there is a prima facie case for their contribution is brought to their attention. Even if it would be really easy for everyone to do something, we cannot expect them to do it if it is not something they know they should do, or know we want them to do.

(6) *Non-harm*. People are more likely to contribute to group actions when they do not perceive their contribution to be harmful. For example, if the secretary of Jackson and Pargetter's procrastinating professor were to accept the book review on his behalf and then the professor were to not complete it, she would have caused more harm (to the author of the book, the journal editors, those who would like to read a review) than had she refused. If she perceives her options in this way, her contribution will be *more* sensitive to what the other person in her group (the professor) will do than it would be if her sole contribution were harmless, or would add incremental value (see (2) above).

(7) *Tipping point: non-contribution.* Depending on the case, sometimes people are most likely to contribute only if they believe that sufficiently many others are likely to *fail* to contribute. That is to say, they will contribute so long as (3) holds, and they believe they are near a non-contribution tipping point. For example, in the last chapter we considered a furniture-removal company. In the modified case with eight members, Kewa may step in only if he believes another member will fail to, because he prefers contributing and the company succeeding to not contributing and the company failing, and he rightly believes he is at a tipping point where his contribution matters.

(8) *Tipping point: contribution.* Again depending on the case, sometimes people are most likely to contribute when they believe that sufficiently many others are *also likely* to contribute, such that there is a reasonable chance that they are at the tipping point, and (3). That is to say, if the production of a good involves a threshold, people will see it as pointless to contribute (unless contributing is independently valuable) unless they

perceive that sufficiently many others are likely to contribute too. For example, imagine that Kewa takes the others out for a big night on the town. He has to decide in the morning whether to show up at work. He can't move furniture alone, so his choice to go to work will be sensitive to his beliefs about whether the others will show up at work or not. Assuming he prefers group success (as in (3)), he will contribute only if he believes his contribution will make that success more likely.

(9) *Difficulty*. I raised the case of 'effort' as a soft constraint in Chapter 5, arguing that the level of effort required for an agent's doing an action that is available to her is irrelevant in establishing what is feasible for her, but that the level of effort required from others for her action to succeed in producing the desired outcome is a soft constraint on its success. For example, Kewa can only lift the piano if Tom, Mark and Jonno help him. Is there a difference in what we expect Tom to do, and therefore the group to do, relative to how hard it would be for him to do it? What if for Tom, lifting the piano would mean coming into work on his day off, extraordinarily hungover, and he would have to try really, really hard to even make it out of bed? Is there some point at which an action is so demanding that we just wouldn't expect a person to do it? The answer seems to be 'yes'. As a condition, 'difficulty' simply notices that people are less likely to do what is extremely difficult or demanding (unless they internalize the project requiring that action in a deep way: certainly people have done many difficult and demanding things in the history of the world). When contributing does not cost individuals too much in terms of effort or resources, especially compared against the perceived gain of the action succeeding, they will be more likely to contribute.

(9) *Informedness*. Lack of knowledge about what others will do, and ungrounded beliefs that they will not contribute, are factors behind (7) and (8). If people satisfy condition (3), then they will be more likely to contribute if their contribution is likely to be crucial. We can manipulate this condition: people are more likely to contribute if they satisfy (3) and believe their contribution to be crucial; so if we want people to contribute, we should make them believe that their contributions are crucial. (The much-discussed case of voter behaviour might seem to undermine this, because here we know that people know their vote won't make a difference, yet they vote anyway. But that is just to say that voting behaviour is not something made more likely by (9) – other cases may be).

These conditions work both separately and in combination to illustrate some of the major conditions affecting the likelihood that a person will contribute to a group

action. As a final thought on this note, some kinds of change and action might be such that they are by their nature unpredictable. Some kinds of political action might only get started by a snowball effect, and that effect might require an irrational first mover, someone who has no reason to expect her action to be efficacious, but who persists nonetheless. Gordon Tullock writes on revolutions that they are often unpredictable because many people outwardly support the status quo, but only because they believe that most others do too. If everyone's belief is conditional in that way, only small changes are required for their beliefs about what others believe to shift, which means that political revolution can be rapid and unprecedented (Tullock, 1971; see also Kuran, 1989; 1991). What this tells us is that even in situations where it looks like people will never be motivated to act in a desired way, things might actually be very different than they look.

8.4 Groups in supergroup contexts

I have said that it doesn't matter much whether we talk about individuals and their group-oriented preferences, or we deal directly in groups and group preferences. And I have given some general conditions under which we should expect individuals to be more likely to contribute to collective actions, and thus conditions under which collective action should be more likely to succeed. But how does this discussion relate to *groups* in *supergroup* contexts, i.e. Germany in the United Nations? Should we predict Germany's actions by reference to the relevant members of that collective, e.g. the power-holders in the German government, and then as a further step predict the United Nations' actions by reference to the preferences of the relevant members of *that* collective, one of which is 'Germany'? And if that's what we should do, does it make sense to think that the conditions making collective action more likely to succeed given in the last section apply to the members of the supergroup, i.e. the groups? Or do they only apply to the members of the groups, i.e. the individuals? I would think that there should be no problem in applying those conditions both to individuals in group contexts, and to groups in supergroup contexts, so long as the groups are rationally interpretable and have the mechanisms available to make informed decisions. If a group can see that it is at a tipping point, and it desires the success of the supergroup- collective action, why not think it will for that reason be more likely to contribute? The main barrier to predicting the actions of groups is their rational interpretability, not their ability to

respond to reasons. So long as states are the kinds of groups who are rationally interpretable, and so long as we have access to some information about their preferences – either directly, or by virtue of having information about the preferences of their individual members – then we should be able to make predictions about how they will act under various of the conditions mentioned in the last section which I claimed make contribution to collective action more likely.

8.5 Conclusion

In this chapter I have tried to at least partially address the question of when collective action is likely to succeed, not conditional upon the collectives' choosing it, which would be appropriate in assessments of 'feasibility-for' collective agents, but unconditionally, as appropriate when we're wondering about the soft constraints on an action's producing an outcome, e.g. for an individual in a group action context, or a group in a supergroup action context. I approached this issue from the angle of groups' and group members' preferences, because we can predict action if we know what people prefer. In Section 8.3.3 I tried to give some general conditions under which a person is likely to do his or her part in a group action, or a group is likely to do its part in a supergroup action. Of course this can at best be only a partial answer, because the details of successful collective action depend on the kind of action at issue, what the group members are like, and what kind of group they are in. We should say very different things about groups of friends than we should say about sports teams, very different things about sports teams than we should say about corporations, and very different things about corporations than we should say about states. To give a better account would be to simply pick one of these and focus more on the details.

Chapter 9

Conclusion

This has been a thesis about the concept of feasibility in political theory. In Chapter 2 I argued that some but not all political theory must be sensitive to feasibility constraints. Only theories which are intended to be directly action-guiding are subject to feasibility constraints. In Chapters 4 & 5 I elaborated two senses of feasibility, the binary sense which plays the role of ruling out recommendations which cannot be realized in practice, and the graded sense which allows for comparisons between alternative sets of recommendations. I argued that the graded sense is necessary in choosing which outcomes to pursue, and takes us further than traditional discussions of 'ought implies can' which have been limited to the ruling-out role. In Chapter 7 I focused in particular on the feasibility of collectives' action, arguing for a particular way of understanding the relation between group and group members' abilities and obligations, so that the feasibility tests formulated in Chapters 4 & 5 extend as well to collective agents.

I have tried to argue that there are two importantly different sets of questions we might ask about feasibility. The first are not indexed to particular agents. Sometimes we want to know whether some outcome, such as ending global poverty, is feasible. We figure that out by asking whether any extant agent has an option that would bring that outcome about. And we figure out how feasible the outcome is by figuring out how likely the most likely action is to bring the outcome about. The second set of questions, in contrast, are indexed to particular agents. Sometimes we want to know whether an agent can bring about a particular outcome that she might be obliged to bring about, such as having contributed some portion of her annual income to global poverty relief. We figure that out by asking what actions are in her option set, focusing on the one most likely to produce the outcome, and asking how likely it is to bring the outcome about assuming that she does it. These two sets of questions exhaust the feasibility assessments relevant to politics. We determine their answers with reference to the hard and soft constraints introduced in Chapters 4 & 5. If hard constraints make it the case than an agent has no action in her option set that could bring about her having contributed to global poverty relief, then that outcome is infeasible for her. If they make it that she does have an (some) action(s) in her option set that could, we ask how likely it is (they are) to bring the outcome about. How likely it is, determined by soft constraints, is what

establishes the comparative feasibility of the outcome.

Interesting questions remain. One is whether the binary test should be dropped altogether in favour of graded feasibility assessments, given the problems raised by strange events with extremely low probability (because almost every action *could* produce almost every outcome); another is whether the conditional probability used in the graded feasibility test should be replaced by a counterfactual test (because problems arise for conditional probabilities that do not arise for counterfactuals with imaging); yet another is how to weigh feasibility considerations against considerations of what is desirable (all I have said is that tradeoffs will be necessary). But these are details. For my own part, I will be happy to move on from talking about what counts as ideal and what counts as non-ideal theory, and about how non-ideal theory should be done (e.g. what kinds of constraints it should satisfy), and to start *doing* non-ideal theory. Having in hand a satisfactory account of what kinds of constraints a theory must satisfy, and what makes a theory most choice-worthy, i.e. maximal feasibility, maximal desirability, and sensitivity to risks, it is possible to identify urgent topics in non-ideal theory and move ahead with addressing them.

Bibliography

Chapter 1

Estlund, David. "Human Nature and the Limits (If Any) of Political Philosophy", paper presented at a meeting of the Canadian Political Science Association, Montreal, June 3rd, 2010.

Gilabert, Pablo. "Feasibility and Global Justice". Manuscript, 2008. Presented at the Workshop on Political Feasibility, at the Australian National University, August, 2008.

Jackson, Frank. & Pargetter, Robert. "Oughts, Options, and Actualism" in *Philosophical Review* (Vol. 95, No. 2, 1986, pp. 233-255).

Räikkä, Juha. "The Feasibility Condition in Political Theory" in *Journal of Political Philosophy* (Vol. 6, No. 1, 1998, pp. 27-40).

Chapter 2

Barry, Christian. & Valentini, Laura. "On Two Critiques of Global Egalitarianism". Manuscript Presented at the Workshop on Political Feasibility, at the Australian National University, August, 2008. (Later published as: Barry, Christian & Valentini, Laura. "Egalitarian Critics of Global Egalitarianism: A Critique" in *Review of International Studies*, 2008).

Brennan, Geoffrey. & Pettit, Philip. "The Feasibility Issue" in Frank Jackson & Michael Smith (Eds.) *Oxford Handbook to Contemporary Philosophy*. Oxford; Oxford University Press, 2005. pp. 258-297.

Brennan, Geoffrey. "Economics" in Robert Goodin & Philip Pettit (Eds.) *A Companion to Contemporary Political Philosophy*. Oxford; Blackwell, 1993.

Buchanan, Allen. *Justice, Legitimacy, and Self-Determination.* Oxford; Oxford University Press, 2004.

Card, Orson Scott. *Ender's Game*. New York; Tom Doherty Associates, 1985.

Chalmers, David. *The Conscious Mind*. Oxford; Oxford University Press, 1996.

Child, Richard. "The dual-component model of justice". Paper presented at the Manchester Metropolitan Workshops in Political Theory, 2–4 September, 2009.

Cohen, Gerald. "Facts and Principles" in *Philosophy & Public Affairs (*Vol. 31, No. 3, 2003, pp. 211-45).

Cohen, G. A. *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press, 2008.

Dworkin, Ronald. S*overeign Virtue: The Theory and Practice of Equality.* Cambridge, MA: Harvard University Press, 2000.

Farrelly, Colin. "Justice in Ideal Theory: A Refutation" in *Political Studies* (Vol. 55, 2007, pp. 844-864).

Gettier, Edmund. "Is Justified True Belief Knowledge?" in *Analysis* (Vol. 23, 1963, pp.121–3).

Gilabert, Pablo. "Feasibility and Global Justice". Manuscript, 2008. Presented at the Workshop on Political Feasibility, at the Australian National University, August, 2008.

Gilabert, Pablo. "Global Justice and Poverty Relief in Nonideal Circumstances," in *Social Theory and Practice* (Vol. 34, 2008, pp. 411-438).

Gilabert, Pablo. & Lawford-Smith, Holly. "Political Feasibility: A Conceptual Exploration" manuscript, 2010.

Gladwell, Malcolm. *Blink: The Power of Thinking Without Thinking*. Boston; Little, Brown & Company, 2005.

Goodin, Robert. "Political Ideals and Political Practice" in *British Journal of Political Science* (Vol. 44, 1995, pp. 635-646).

Goodin, Robert. "The Bioethics of Second Best" in Joseph Millum and Ezekiel Emanuel (Eds.) *Global Justice and Bioethics*, forthcoming.

Intersex Society of North America website http://www.isna.org/faq/frequency accessed 18/08/10.

Jackson, Frank. "Epiphenomenal Qualia" in *Philosophical Quarterly* (Vol. 32, 1982, pp. 127–36.

Jackson, Frank. "Narrow Content and Representation – or Twin Earth Revisited" in *Proceedings and Addresses of the American Philosophical Association* (Vol. 77, No. 2, 2003, pp. 55-70).

Korsgaard, Christine. "Realism and constructivism in twentieth-century moral philosophy" in *Philosophy in America at the Turn of the Century: APA Centennial Supplement to Journal of Philosophical Research*. Charlottesville, VA: Philosophy Documentation Center, 2003, pp. 99–122. Available at 〈www.pdcnet.org/pdf/8Korsgaard.pdf〉 .

Lawford-Smith, Holly. 'Ideal Theory: A Reply to Valentini.' *Journal of Political Philosophy* (Vol. 18, No. 3, 2010, pp. 357-368).

Lewis, David. *Counterfactuals*. Oxford; Blackwell, 1973.

Lipsey, R.G. & Lancaster, Kevin. "The General Theory of the Second Best" in *The Review of Economic Studies* (Vol. 24, No. 1, 1956-7, pp. 11-32).

Miller, David. "Political Philosophy for Earthlings" in David Leopold & Marc Stears (Eds.) *Political Theory: Methods and Approaches.* Oxford; Oxford University Press, 2008, pp. 29-48.

Mills, Charles. ""Ideal Theory" as Ideology"" in *Hypatia* (Vol. 20, No. 3, 2005, pp. 165-184).

Murphy, Liam. *Moral Demands in Nonideal Theory*. Oxford: Oxford University Press, 2000.

Okin, Susan Moller. *Justice, gender, and the family*. New York, Basic Books, 1989.

O'Neill, Onora. "Abstraction, Idealization and Ideology in Ethics" in J.D.G Evans (Ed.) *Moral Philosophy and Contemporary Problems.* Cambridge; Cambridge University Press, 1987.

O'Neill, Onora. *Bounds of Justice*. Cambridge: Cambridge University Press, 2000.

Pogge, Thomas. "Cohen to the Rescue!" in *Ratio* (Vol. XXI, Dec., 2008, pp. 454-475).

Rawls, John. *A Theory of Justice*. Cambridge, MA: Harvard University Press, 1971.

Sen, Amartya. "What Do We Want from a Theory of Justice?" in *Journal of Philosophy* (Vol. 103, 2006, pp. 215–38).

Simmons, John. "Ideal and Non-Ideal Theory" in *Philosophy & Public Affairs* (Vol. 38, No. 1, 2010, p. 5-36).

Stemplowska, Zofia. "What's ideal about ideal theory" in *Social Theory and Practice* (Vol. 34, 2008, pp. 319-340).

Swift, Adam. "The Value of Philosophy in Non-Ideal Circumstances" in *Social Theory and Practice* (Vol. 34, No. 3, 2008, pp. 363-387).

Valentini, Laura. "On the apparent paradox of ideal theory" in. *Journal of Political Philosophy* (Vol. 17, 2009, pp. 332–55).

Zuolo, Federico. "Bridging the gap between ideal and nonideal theory: the concept of self-efficacy in normative political theories." Paper presented at the Manchester Metropolitan Workshops in Political Theory, 2–4 September 2009.

Chapter 3

Barry, Christian. & Valentini, Laura. "On Two Critiques of Global Egalitarianism". Manuscript, 2008. Presented at the Workshop on Political Feasibility, at the Australian National University, August, 2008. (Later published as: Barry, Christian & Valentini, Laura. "Egalitarian Critics of Global Egalitarianism: A Critique" in *Review of International Studies*, 2008).

Brennan, Geoffrey. & Southwood, Nicholas. "Feasibility in Action and Attitude" in *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz.* Ed. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson, 2007. Available online at www.fil.lu.se/hommageawlodek.

Brennan, Geoffrey. & Pettit, Philip. "The Feasibility Issue," in Frank Jackson and Michael Smith (Eds.) *The Oxford Handbook of Contemporary Philosophy.* Oxford; Oxford University Press, 2005, pp. 258-79.

Brock, Gillian. *Global Justice: A Cosmopolitan Account*. Oxford; Oxford University Press, 2009.

Buchanan, Allen. *Justice, Legitimacy, and Self-Determination.* Oxford; Oxford University Press, 2004.

Carnap, Rudolph. *Logical Foundations of Probability*. Chicago; University of Chicago Press, 1950.

Cohen, Gerald. *Why Not Socialism?* Princeton: Princeton University Press, 2009.

Cohen, Gerald. "Why Not Socialism?" in E. Broadbent (Ed.) *Democratic Equality. What Went Wrong?* Toronto; University of Toronto Press, 2001, pp. 58-78.

Cranston, Maurice. "Human Rights, Real and Supposed" in P. Hayden (Ed.) *The Philosophy of Human Rights*. St. Paul; Paragon House, 2001. pp. 163-173.

Dennett, Daniel. *Darwin's Dangerous Idea*. New York; Touchstone, 1995.

Gates, Bill. "Why We Need Innovation, Not Just Insulation" in *The Gates Notes* (24th Jan, 2010) available online at http://www.thegatesnotes.com/Thinking/article.aspx?ID=47.

Gilabert, Pablo. "The feasibility of basic socioeconomic human rights: A conceptual exploration" in *Philosophical Quarterly* (Vol. 59, No. 237, Oct. 2009, pp. 659-681).

Gilabert, Pablo. "Feasibility and Socialism" in Journal of Political Philosophy, forthcoming.

Hawthorn, Geoffrey. *Plausible Worlds*. Cambridge; Cambridge University Press, 1991.

Held, David. "Democracy: from city-states to a cosmopolitan order?" in *Political Studies* (Vol. 40, Aug., 1992, pp. 10-39).

Hirschbiegel, Oliver. [Movie] "Der Untergang" (2004).

Jensen, Mark. "The Limits of Practical Possibility" in *Journal of Political Philosophy* (Vol. 17, No. 2, 2009, pp. 168-184).

Lawford-Smith, Holly. "Cosmopolitan Global Justice: Brock v. the Feasibility Sceptic" forthcoming in *Global justice: Theory Practice Rhetoric.*

Lewis, David. "Counterfactual Dependence and Time's Arrow" in *Noûs* (Vol. 13, No. 4, Nov. 1979, pp. 455-476).

Nozick, Robert. *Anarchy, State and Utopia*. New York; Basic Books,1974.

Pogge, Thomas. "Cosmopolitanism and sovereignty" in *Ethics* (Vol. 103, No. 1, 1992, pp. 48-75).

Pettit, Philip. "Is Criminal Justice Politically Feasible?" in *Buffalo Criminal Law Review* (Vol. 5, 2002, p. 427-450).

Rawls, John. *A Theory of Justice*. Cambridge, MA; Harvard University Press, 1971.

Raz, Joseph. *The Morality of Freedom*. Oxford; Oxford University Press, 1986.

Tetlock, Philip. Lebow, Richard. & Parker, Noel (Eds). *Unmaking the West: "What If?" Scenarios that Rewrite World History*. Ann Arbor; University of Michigan Press, 2006.

Wenar, Leif. "Why Rawls is Not a Cosmopolitan Egalitarian" in R. Martin & D. Reidy (Eds.) Rawls's Law of People's: A Realistic Utopia? Oxford; Blackwell, 2006, pp. 95-113.

Wright, Erik Olin. *Envisioning Real Utopias*. London; Verso, 2010.

Chapter 4

Bird, Alexander. "Dispositions and Antidotes" in *The Philosophical Quarterly* (Vol. 48, 1998, pp. 227–234).

Blackburn, Simon. *Essays in Quasi-Realism.* Oxford; Oxford University Press, 1993.

Bowles, Samuel. & Gintis, Herbert. "The Origins of Human Cooperation" in Peter Hammerstein (Ed). *The Genetic and Cultural Origins of Cooperation*. Cambridge; MIT Press, 2003.

Boyd, Robert. & Richerson, Peter. "Solving the Puzzle of Human Cooperation" in S. Levinson (Ed.) *Evolution and Culture*. Cambridge; MIT Press, 2005, pp. 105-132.

Brennan, Geoffrey. & Pettit, Philip. "The Feasibility Issue" in Frank Jackson & Michael Smith (Eds.) *Oxford Handbook to Contemporary Philosophy*. Oxford; Oxford University Press, 2005. Pp. 258-297.

Broome, John. *How to be Rational*. Oxford; Mimeo, forthcoming.

Broome, John. *Weighing Goods: Equality, Uncertainty, and Time*. Oxford; Blackwell, 1991.

Brink, David. *Moral realism and the foundations of ethics.* New York; Cambridge University Press, 1989.

Chalmers, David. "Does Conceivability Entail Possibility?" in T. Gendler & J. Hawthorne, Eds. *Conceivability and Possibility*. Oxford; Oxford University Press, 2002, pp.145-200.

Cohen, Gerald. "Facts and Principles" in *Philosophy & Public Affairs* (Vol. 31, No. 3, 2003, pp. 211-45).

Collingridge, David. "'Ought-implies-can' and Hume's rule" in *Philosophy* (Vol. 52, 1977, p. 351).

Crisp, Roger. *Reasons and the Good*. Oxford; Clarendon Press, 2006.

Dancy, Jonathan. *Moral Reasons*. Oxford; Blackwell, 1993.

Dawkins, Richard. *The Blind Watchmaker*. United States; Norton, 1986.

Dennett, Daniel. *Darwin's Dangerous Idea*. New York; Touchstone, 1995.

Dretske, Fred. *Knowledge and the flow of Information.* Massachusetts; MIT Press, 1983.

Fara, Michael. "Dispositions" in *Stanford Encyclopaedia of Philosophy*, available online at http://plato.stanford.edu/entries/dispositions/ (26th Jul. 2006). Accessed 18/08/10.

Fara, Michael. "Dispositions and Habituals" in *Noûs* (Vol. 39, 2005, pp. 43–82).

Fara, Michael. "Masked Abilities and Compatibilism" in *Mind* (Vol. 117, 2008, pp. 843–865).

Frankfurt, Harry. "Alternate Possibilities and Moral Responsibility" in *The Journal of Philosophy* (Vol. 66, 1969, pp. 829–839).

Gilabert, Pablo. "Feasibility and Global Justice". Presented at the Workshop on Political Feasibility, Australian National University, August 2008.

Grice, Paul. *Studies in the Ways of Words.* Cambridge, MA; Harvard University Press, 1989.

Harrison, Jonathan. "Ethical naturalism" in P. Edwards (Ed.). *The encyclopedia of philosophy,* Vol. 3. New York; Macmillan, 1967, pp. 69-71.

Hawthorne, John. "Chance and Counterfactuals" in *Philosophy and Phenomenological Research* (Vol. LXX, No. 2, Mar. 2005, pp. 396-405).

Hintikka, Jaako. *Knowledge and Belief.* Ithaca, New York; Cornell University Press, 1962.

Jackson, Frank. *From Metaphysics to Ethics.* Oxford; Oxford University Press, 1998.

Jensen, Mark. "The Limits of Practical Possibility" in *Journal of Political Philosophy* (Vol. 17, No. 2, 2009, pp. 168-184).

Johnston, M. "How to Speak of the Colours" in *Philosophical Studies* (Vol. 68, 1992, pp. 221–263).

Joyce, Richard. *The Evolution of Morality*. Massachusetts; MIT Press, 2006.

Kripke, S. *Naming and Necessity*. Cambridge, Mass.; Harvard University Press, 1980.

Lewis, David. "The Paradoxes of Time Travel" in *American Philosophical Quarterly*, (Vol. 13, 1976, p. 6).

Maier, John. "Abilities" in *Stanford Encyclopaedia of Philosophy*, available online at http://plato.stanford.edu/entries/abilities/ (26th Jan. 2010). Accessed 17/08/10.

Martin, C. B. "Dispositions and Conditionals" in *The Philosophical Quarterly* (Vol. 44, 1994, pp. 1–8).

Moore, G.E. *Principia Ethica.* Cambridge; Cambridge University Press, 1903.

Parfit, Derek. *On What Matters*. Oxford; Oxford University Press, forthcoming.

Pigden, Charles. "Ought-implies-can: Erasmus Luther and R. M. Hare" in *Sophia* (Vol. 29, No. 1, April, 1990). pp. 2-30.

Pigden, Charles. *The Reluctant Nihilist* (draft book distributed for teaching purposes). Dunedin, University of Otago, 1991.

Pigden, Charles. "Nihilism, Nietzsche and the Doppelganger Problem" in *Ethical Theory and Moral Practice* (Vol. 10, No. 5, 2007, pp. 441-456).

Pigden, Charles. "Introduction" (pp. 1-29) and "If not Non-Cognitivism then What?" (pp. 80-104) in Charles Pigden (Ed.) *Hume on Motivation and Virtue*. Houndmills; Palgrave Macmillan, 2009.

Popper, Karl. *Conjectures and Refutations.* New York; Routledge & Kegan Paul, [1963] 2004.

Prior, Arthur. "The autonomy of ethics" in *Australasian Journal of Philosophy* (Vol. 38, 1960, pp. 197-206).

Ryle, Gilbert. *The Concept of Mind*, London; Hutchinson, 1949.

Rynin, David. "The autonomy of morals" in *Mind* (Vol. 66, 1957, pp. 308-317).

Rosenberg, Alex. & McShea, Daniel. *Philosophy of Biology: A Contemporary Introduction.* New York; Routledge, 2008.

Sapontzis, Steve. " 'Ought' does imply 'can' " in *The Southern Journal of Philosophy* (Vol. 29, 1991, pp. 383-393.

Schurz, Gerhard. "How far can Hume's is-ought thesis be generalized? An investigation in alethic-deontic modal predicate logic" in *Journal of Philosophical Logic* (Vol. 20, 1991, pp. 37-95).

Searle, John. "How to derive "ought" from "is"" in *The Philosophical Review* (Vol. 73, 1964, pp. 43-58).

Sinnott-Armstrong, Walter. " 'Ought' Conversationally Implies 'Can' " in *Philosophical Review* (Vol. 93, 1984, pp. 249-261).

Sidgewick, Henry. *The Methods of Ethics*. London; MacMillan, 1874.

Smith, A. D. "Dispositional Properties" in *Mind* (Vol. 86, 1977, pp. 439–445).

Stalnaker, Robert. "Common Ground" in *Linguistics and Philosophy* (Vol. 25, 2002, pp. 701-721).

Stalnaker, Robert. "Presuppositions" in *Journal of Philosophical Logic* (Vol. 2, 1973, pp. 447-457).

Streumer, Bart. "Does 'Ought' Conversationally Implicate 'Can'?" in *European Journal of Philosophy* (Vol. 11, No. 2, 2003, p. 219-228).

Vranas, Peter. "I Ought Therefore I Can" in *Philosophical Studies* (Vol. 136, No. 2, 2007, pp. 167-216).

Ward Smith, James. "Impossibility and Morals" in *Mind* (Vol. LXX, No. 279, 1961, p. 374).

Wedgewood, Ralph. *The Nature of Normativity*. Oxford; Oxford University Press, 2007.

Chapter 5

Brennan, Geoffrey. & Southwood, Nicholas. "Feasibility in Action and Attitude" in *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*. Ed. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson, 2007. Available online at www.fil.lu.se/hommageawlodek.

Estlund, David. "Human Nature and the Limits (If Any) of Political Philosophy", paper presented at a meeting of the Canadian Political Science Association, June 3rd, 2010.

Hájek, Alan. "Counterfactual reasoning (philosophical aspects) – quantitative" in N.J Smelser & P.B Baltes, *International Encyclopedia of the Social and Behavioural Sciences*. Oxford, Elsevier, 2002, pp. 2872-2874.

Hochschild, Adam. *King Leopold's Ghost*. New York; Mariner Books, 1999.

Jackson, Frank. & Pargetter, Robert. "Oughts, Options, and Actualism" in *Philosophical Review* (Vol. 95, No. 2, 1986, pp. 233-255).

Maier, John. "Abilities" in *Stanford Encyclopaedia of Philosophy*, available online at http://plato.stanford.edu/entries/abilities/ (26th Jan. 2010). Accessed 17/08/10.

Singer, Peter. "Famine, Affluence, and Morality" in *Philosophy and Public Affairs* (Vol. 1, 1972, pp. 229-243).

Singer, Peter. *The Life You Can Save*. Melbourne; Text Publishing, 2009.

Woodard, Christopher. "Group-Based Reasons for Action" in Ethical Theory and Moral Practice (Vol. 6, 2003, pp. 215-229).

Chapter 6

Brennan, Geoffrey. & Southwood, Nicholas. "Feasibility in Action and Attitude" in *Hommage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*. Ed. T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson, 2007. Available online at www.fil.lu.se/hommageawlodek.

Byrne, R. M. J. *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA; MIT Press, 2005.

Fine, Kit. "Review of Lewis's *Counterfactuals*" in *Mind* (Vol. 84, 1975, pp. 451-458).

Hawthorne, John. "Chance and Counterfactuals" in *Philosophy and Phenomenological Research* (Vol. LXX, No. 2, Mar. 2005, pp. 396-405).

Horgan, Terry. "Counterfactuals and Newcomb's Problem" in R. Campbell and L. Sowden eds., *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem.* Vancouver, BC; University of British Columbia Press, 1985.

Joyce, James. *The Foundations of Causal Decision Theory.* Cambridge; Cambridge University Press, 1999.

Kahneman, D. & Tversky, A. "The simulation heuristic," in D. Kahneman, P. Slovic, and A. Tversky (Eds.) Judgement under Uncertainty. Cambridge: Cambridge University Press, 1982.

Kyberg, Henry. *Probability and the Logic of Rational Belief*. Middletown, Connecticut; Wesleyan University Press, 1961.

Lewis, David. "Counterfactual Dependence and Time's Arrow" in *Noûs* (Vol. 13, No. 4, Nov. 1979, pp. 455-476).

Lewis, David. *Counterfactuals.* Cambridge, MA; Harvard University Press, 1973.

List, Christian. & Pettit, Philip. "On the Many as One" in *Philosophy and Public Affairs* (Vol. 33, 2005, pp. 377-390).

List, Christain. & Deitrich, Franz. "Judgement aggregation under constraints," in T. Boylan and R. Gekker (Eds.). *Economics, Rational Choice, and Normative Philosophy.* London; Routledge, 2008.

Price, Huw. "Against Causal Decision Theory" in *Synthese* (Vol. 67, 1986, pp. 195–212).

Roese, N. J. & Olson, J. "The structure of counterfactual thought" in *Personality and Social Psychology Bulletin* (Vol. 19, 1993, pp. 312–19).

Roese, N. J. & Olson, J. "Functions of counterfactual thinking," in N. J. Roese and J. M.

Olson (Eds.) *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Mahwah, NJ; Erlbaum, 1995.

Skyrms, Brian. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws.* New Haven, CT; Yale University Press, 1980.

Sobel, Jordan Howard. *Taking Chances: Essays on Rational Choice.* Cambridge; Cambridge University Press, 1994.

Williamson, Timothy. *The Philosophy of Philosophy.* Oxford; Blackwell, 2007.

Chapter 7

Bratman, Michael. "Shared Cooperative Activity" in *The Philosophical Review* (Vol. 101, No. 2, Apr., 1992, pp. 327-341).

Hume, David. *A Treatise of Human Nature*. 2nd Edition. (Ed. L.A Selby-Bigge). Oxford; Clarendon Press, 1978.

Jackson, Frank. & Pargetter, Robert. "Oughts, Options, and Actualism" in *Philosophical Review* (Vol. 95, No. 2, 1986, pp. 233-255).

Lewis, David. "Semantic Analyses for Dyadic Deontic Logic" in *Papers in Ethics and Social Philosophy*. Cambridge; Cambridge University Press, 2000, pp. 5-19.

Lewis, David. *Counterfactuals*. Harvard; Blackwell, 1973, pp. 79-80.

Pettit, Philip. & Schweikard, David. "Joint Actions and Group Agents" in *Philosophy of the Social Sciences* (Vol. 36, 2006, pp. 18-39).

Tuomela, Raimo. "We Will Do It: An Analysis of Group Intentions" in *Philosophy and Phenomenological Research* (Vol. 51, No. 2, Jun. 1991, pp. 249-277).

Chapter 8

Arrow, Kenneth. "A Difficulty in the Concept of Social Welfare" in *Journal of Political Economy* (Vol. 58, No. 4, August, 1950, pp. 328-346).

Arrow, Kenneth. *Social Choice and Individual Values*. 1st Ed. Connecticut, Yale University Press, 1951.

Arrow, Kenneth. *Social Choice and Individual Values*. 2nd Ed. Connecticut, Yale University Press; 1963.

Blackburn, Simon. "Game Theory and Rational Choice" in *Ruling Passions.* Oxford; Oxford University Press, 1998.

Bowles, Samuel. & Gintis, Herbert. "Social Preferences, *Homo Economicus,* and *Zoon Politikon*" in Robert Goodin & Charles Tilly (Eds). *The Oxford Handbook of Contextual Political Analysis*. London; Oxford, 2006.

Dreier, James. "Decision Theory and Morality" in Alfred Mele (Ed.) *The Oxford Handbook of Rationality*. Oxford; Oxford University Press, 2003.

Fishburn, Peter. "Utility and Subjective Probability" in *Handbook of Game Theory with Economic Applications*, Vol. 2. Amsterdam; Elsevier, 1994.

French, Peter A. "The Corporation as a Moral Person" in *American Philosophical Quarterly* (Vol. 16, No. 3, Jul. 1979, pp. 207-215).

Gilbert, Margaret. "Collective Preferences, Obligations and Rational Choice" in *Economics and Philosophy* (Vol. 17, 2001, pp. 109-119).

Gilbert, Margaret. "Modeling Collective Belief" in *Synthese* (Vol. 71, No. 1, Oct. 1987, pp. 185-204).

Hardin, Russell. *Collective Action*. London; John Hopkins University Press, 1982.

Joyce, James. *The Foundations of Causal Decision Theory*. Cambridge; Cambridge University Press, 1999.

Kiewiet, D. Roderick. *Micropolitics and macroeconomics*. Chicago; Chicago University Press, 1983.

Kinder, Donald. & Kiewiet, D. Roderick. "Sociotropic politics: the American case" in *British Journal of Political Science* (Vol. 11, 1981, pp. 129-161).

Kuran, Timur. "Sparks and Prarie Fires: A Theory of Unanticipated Political Revolution" in *Public Choice* (Vol. 61, 1989, pp. 41-74).

Kuran, Timur. "Private and Public Preferences" in *Economics and Philosophy* (Vol. 6, 1990, pp. 1-26).

Kuran, Timur. "Now or Never: the element of surprise in the East European revolution of 1989" in *World Politics* (Vol. 44, No. 1, October, 1991, pp. 7-48).

Ledyard, John. "Public Goods: A Survey of Experimental Research" in Roth & Kagel, *The Handbook of Experimental Economics.* New Jersey; Princeton, 1995.

List, Christian. & Pettit, Philip. "Aggregating sets of judgments: an impossibility result" in *Economics and philosophy* (Vol. 18, No. 1, 2002. pp. 89-110).

List, Christian. & Pettit, Philip. "Aggregating sets of judgements: two impossibility results compared" in *Synthese* (Vol. 140, No. 1-2, 2004, pp. 207-235).

Olson, Mancur. *The Logic of Collective Action: Public Goods and the Theory of Groups.* Harvard; Harvard University Press, [1965] 1971.

Ramsey, Frank. "Truth and Probability" in R. Braithwaite (Ed.) *The Foundations of Mathematics and other Logical Essays*. London; Kegan Paul, 1931, pp. 156-198.

Resnik, Michael. *Choices: An Introduction to Decision Theory*. Minneapolis; University of Minnesota Press, 1987.

Rohrschneider, R. "Citizens' attitudes toward environmental issues: selfish or selfless?" in *Comparative Political Studies* (Vol, 21, 1988, pp. 347-367).
Ross, Don. "Game Theory" in *Stanford Encyclopedia Online* ([Jan. 1997] March 2006).

Savage, Leonard. *The Foundations of Statistics*. 2nd Ed. New York; Dover Press, [1954] 1972.

Sudgen, Robert. "Team Preferences" in *Economics and Philosophy* (Vol. 16, 2000, pp. 175-204).

Sunstein, Cass. "True Lies" at *The New Republic (*Dec. 25th, 1995). Available online at http://www.tnr.com/article/books-and-arts/true-lies accessed 14/09/10).

Tullock, Gordon. "The Paradox of Revolution" in *Public Choice* (Vol. 11, No. 1, 1971, pp. 89-99).

Weatherson, Brian. *Introduction to Decision Theory.* Manuscript. 2010

Wendt, Alexander. "The state as a person in international theory" in *Review of International Studies* (Vol. 30, 2004, pp. 289-316).