The concept of feasibility and its role in moral and political philosophy

Daniel Guillery

UCL

Thesis to be submitted for the degree of MPhil Stud

I, Daniel Guillery, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed:

Abstract

This thesis concerns the nature of the concept of feasibility and its role in constraining moral and political philosophy: to what extent and in what way facts about feasibility ought to constrain what moral and political theory say.

I begin in my first chapter by giving an account of feasibility, that is, by attempting to understand what we mean when we say that some outcome is or isn't feasible. I argue against the various attempts that have been made in the literature to give a binary definition (e.g. Gilabert and Lawford-Smith, Räikkä). There is a multiplicity of different possible sharpenings of the term 'feasible', no single one of which is obviously privileged. Different sharpenings hold fixed different ranges of facts, making different sets of proposals feasible.

In the remainder of the thesis, I go on to relate this account of feasibility to moral and political theory. I argue that it is not clear which sharpenings of 'feasibility' constrain which sorts of moral theory. I engage with the literature on 'ideal theory', arguing that theory constrained only by expansive (permissive) sharpenings of 'feasibility' (which is one thing that could be meant by 'ideal theory') is useful and important for the purpose of practical action guidance.

I thus draw two important conclusions. The first is the thesis of the first chapter about the concept of feasibility. I then build on this to get to a more substantive methodological conclusion, that theory constrained only by permissive (unrealistic) feasibility constraints is useful.

Table of Contents

Introduction

It is common in political philosophy or in the practice of politics for a theory or a practical proposal to be criticised or rejected for not being feasible. A *feasibility critique* says of some normative political theory or practical proposal either that it is mistaken or that it is uninteresting or unimportant because the observance of its requirements is unfeasible. The importance of such critiques is evident in the domain of real politics. Here, it is rarely questioned that these are good grounds for the rejection of proposals; rather debates tend to centre on whether or not it is true that the proposal is unfeasible. Such critiques are similarly important in the domain of political philosophy. Here too, it is often thought that if the observance of some principle is unfeasible, then it cannot form a part of a correct (or interesting) moral (or political) theory. If this is not always made explicit it is often tacitly assumed. Such critiques have taken on additional prominence with the development of objections to what has been called 'ideal theory'. One important criticism of such theory has been that it offers recommendations or requirements that are not feasible.

It is easy to see that feasibility is a widespread consideration in political philosophy. Examples abound of theories or proposals that are criticised (and rejected) for being unfeasible. Take, for example, the model of participatory democracy. It calls for widespread (or universal) participation of the members of a society in decision-making. This model has been rejected by many political theorists because it is not thought to be feasible. John Stuart Mill, for example, thought that we should reject this theory: 'since all cannot, in a community exceeding a single small town, participate personally in any but some very minor portion of the public business, it follows that the ideal type of a perfect

government must be representative' (a model where citizens do not participate directly in decision-making).[1]

Another political theory that is frequently objected to on grounds of feasibility is anarcho-communism: the political theory that calls for something like distribution according to need, or equal distribution, alongside the absence of a coercive state. Arguably *the* most common objection to this position is that though what it calls for might be desirable, it is not achievable. David Miller, for example, argues that 'a central agency seems necessary to maintain any society-wide distribution of resources'.[2] Therefore, it is not feasible to achieve the distribution required without a state.

More mainstream political theories are also criticised for being unfeasible. For example, one criticism of luck egalitarianism (the doctrine that says that inequalities are only just when they reflect differences in choices for which we can be held responsible) is that it is unfeasible to determine what is due to choice and what not, and thus that it is unfeasible to observe the luck-egalitarian principle.[3] A principle of equality of opportunity is also open to feasibility critiques, since it might be thought, for instance, that it is prevented from being feasible by the strength of the institution of the family, which perhaps leads inevitably to certain inequalities in life chances.[4]

Given the importance of feasibility critiques in political theory, then, it seems important to have an idea of what it means to say that some proposal is or is not feasible and what significance it has. What do we mean when we say that participatory democracy, or anarcho-communism, is not feasible? How do we adjudicate these claims? What are their truth conditions? We want to know when unfeasibility warrants the rejection of some theory or proposal, either per se, or as a theory that is interesting or useful. The first step

[1] Mill (1861) 234
[2] Miller (1984) 172
[3] Roemer (2002) aims to defend against such a critique.
[4] Rawls (1999a) makes this point, pg. 74.

towards answering this will of course be to get an idea of what we mean by 'unfeasibility'. For example, then, is an electoral system of proportional representation (PR) feasible in the UK? To answer this, we will need an idea of what we mean by 'feasible'. It is by no means immediately clear. In a sense it certainly is feasible: it does not seem to contravene anything deep or basic about UK citizens or humans generally; we could live under a system of PR without dramatically changing our natures. On the other hand, though, given the current electoral system and the way legislators are selected, it might be thought that it is not feasible, since the first-past-the-post system tends to give the majority of parliamentary seats to the big parties, who do not have an interest in changing the electoral system. However we decide the question of what 'feasible' means, another question arises, which is whether a theory recommending PR is to be rejected (on grounds of feasibility). In order to answer these questions for any proposal: whether it is feasible and if it is not, what this means for the proposal, we will need an account of what 'feasibility' means and of how facts about feasibility relate to moral and political theories or proposals. Recently, some philosophers, though few, have started to address these questions, and that is also what I seek to do in this thesis.

Desirability and feasibility

Before I proceed to give an account of the concept of feasibility, it is important to note that feasibility critiques are frequently not as simple as the above description suggested. Often when a proposal or principle is criticised for being unfeasible we do not really mean that it is unfeasible simpliciter. Often these critiques are mixed up with questions of desirability, that is, with evaluative or normative questions. The concept of feasibility, as I understand it, is not an evaluative or normative one. However, when we say that some proposal is not feasible, often we mean that it is not feasible in conjunction with certain other things that we take to be more desirable or with the observance of other principles,

which we take to be weightier than the proposal in question. A feasibility critique of this form, then, says something like 'given that we should do *x* or realise (values *v*, proposals *p*, principles *q*), it is not feasible to do/bring about/realise *y*'. For example, feasibility critiques of anarchism are often not as simple as the objection to anarcho-communism discussed above. The most common way of objecting to anarchism is to say that it is not feasible in conjunction with the realisation of certain weighty values. One thought is that it is not feasible to achieve a stateless society alongside a certain level of personal security or peace.[5] Another is that it is not feasible together with distributive justice.[6] Sometimes such objections are simply put by saying that anarchism is not feasible or not possible. However, this is usually not what is meant. It is not really thought that it is just not feasible to bring about a stateless society. Rather, it is thought that if we did so, we would be morally worse off than we actually are; the desirability of distributive justice or peace or whatever is taken for granted. It is not feasible to achieve a stateless society in a desirable way.

The fact that feasibility critiques are often mixed up with questions of desirability does not mean that questions of feasibility themselves are evaluative or normative. The feasibility question is separate from the desirability question.[7] Or rather, the two *can* be separated, though often we put them together. In general, what tends to be most important to know is not just whether some proposal is feasible, but rather whether it is feasible in an all-things-considered desirable way. We need to ask whether it is feasible in conjunction with the realisation of those other principles or values that would make it all-things-considered desirable.

[5] Miller (1984) argues that it is not feasible to control antisocial behaviour without formal laws (173-9). Landauer (1959) argues that 'such evidence as we have does not indicate that ill intentions will cease to exist if repressive force disappears' (128).
[6] Miller (1984) makes this argument (172-3), as, in effect, does Wolff (1996).
[7] Gilabert (2008, 415), Gilabert and Lawford-Smith (2012, 816-7) and Wiens (forthcoming, 9) also make this point. Gilabert (2008) thus says that 'normative political argument looks for the *intersection* between desirability and feasibility' (415).

Outline

My thesis will argue for two key claims, though the argument for one will depend on the other. The first is about the meaning of 'feasibility': it is the claim that there is no single privileged binary definition of the term available, but rather a multitude of different possible sharpenings. I will argue for this in my first chapter. I will give a general scalar definition in terms of binary definitions for the different sharpenings. I will build on this account of feasibility through the second and third chapters to work towards my second key claim, which is that 'unrealistic' political philosophy (that is, theory that is not constrained by restrictive sharpenings of 'feasibility') is worthwhile and, more specifically, useful for guiding action in the real world.

In the second chapter, I will argue that Norman Daniels's distinction between achievability and sustainability does not provide the key to the difference between cases in which feasibility considerations warrant the rejection of proposals (either as correct or as interesting) and those in which they do not. However, I claim, given the conclusions of the first chapter, it is not clear on exactly which sharpenings of 'feasibility' different sorts of 'ought'-claims must be feasible in order to be correct or interesting. There are, though, conceivable (though not necessarily interesting) types of theory that could be carried out for each possible sharpening. Which of these provide *correct* moral principles is not necessarily the same question as the question on which sharpenings a theory must be feasible in order to be *worthwhile*. I focus on the latter in this thesis, though I do not attempt to provide any sort of complete answer to the question.

In the third chapter I come to my defence of the importance of 'unrealistic' theory. Though the terminology is contested, this is one way we might understand what is meant by 'ideal theory', so my defence of this sort of political philosophy could be understood as a contribution to the ideal/non-ideal theory debate. Since, given my first claim, such

theory cannot simply be rejected by saying what it recommends is unfeasible (since it is feasible on some sharpenings), a critic would need to argue that only certain sharpenings of 'feasibility' should constrain the sort of moral and political theory that we should be interested in. I reject certain possible arguments to this effect. Theorising with expansive feasibility constraints is important and relevant to deciding what immediate short-term actions to take.

Chapter 1

Feasibility constraints

Feasibility can be understood in many ways. To take anarchism as an example, one understanding of what it would mean to say that anarchism is not feasible is that given states as they currently are, human motivation as it currently is, and the current political situation (and the number and influence of anarchists) and so on, states are not going to stop imposing their laws on people. At the opposite extreme, the claim could be understood as the claim that anarchism is not physically, or metaphysically, or even logically, possible. Somewhere in between is the claim that a stateless society (or whatever anarchism is taken to require) is made impossible by certain facts about human nature, even if we allow things like the current political situation and current preferences to change. It is not obvious that any one of these things is what 'feasibility' standardly means or should properly mean. Sometimes in a discussion it is clear which constraints we are accusing a proposal of violating when we say it is not feasible. Often, though, it is not.

These different things we could mean point to different possible ways of sharpening the term 'feasibility'. There is not one single binary or categorical conception of feasibility.[8] Rather, there is a whole range of possible sharpenings of the term 'feasible'. Below, I will argue that though when we sharpen 'feasible' in any one of these ways it may become a binary matter whether or not some proposal is feasible, the only general definition of 'feasibility' that can be given is comparative. This range of sharpenings can be understood

[8] Gilabert (2008) is aware of this imprecision, noting that intuitions pull in both ways about how expansive the definition of 'feasibility' should be and that there is no obvious way to achieve a balance (415-6).

using the notion of feasibility constraints (FCs).[9] An FC is a set of facts about the world that are held fixed and a set of facts that are allowed to vary. Thus, the different understandings of the feasibility of anarchism above involve different FCs. One FC holds fixed the current political situation etc. while others do not. As we vary more and more facts as well as facts that are more and more hard to change, we progress up a feasibility constraint scale. Thus, the lowest FC is one where no facts are allowed to vary, everything is held constant. The only thing that will come out as feasible on this FC is the status quo. At the other end, presumably the highest possible FC is one that allows all facts to vary. On this everything will come out as feasible (perhaps excluding the logically impossible). There will obviously, though, be a large range of FCs in between the lowest and the highest extremes, as we allow more and more things to vary. Low FCs are more realistic and restrictive, while high FCs are less realistic and more expansive. (When referring to the position of FCs on this scale I will use 'high' and 'low' in this way, though they could equally be used the opposite way around. A *low* FC holds fixed a larger range of facts and so is more restrictive: fewer things are feasible, while a *high* FC holds fixed less and so is more permissive). When we ask whether some proposal is feasible on some FC we ask whether it is feasible allowing the chosen range of facts to vary and holding all the others constant. I will attempt to give a definition of what this means in later sections of this chapter.

Now, the FC-scale will not in fact be quite as simple as the picture I have just painted suggests. We do not just progress up the scale simply by adding 'the next' most changeable fact to the list of facts allowed to change. The scale is not linear in this way.

[9] Cf. the discussion of feasibility frontiers in Hamlin and Stemplowska (2012). They recognise a variety of ways in which feasibility can be defined, which they represent as a range of feasibility frontiers (52ff).

One could theoretically choose almost any set of facts and allow these to vary.[10] However, though the scale is not linear in this sense, it does assume that these disparate FCs can be ranked in terms of 'realisticness'. The first thing to note before we can get an idea of how FCs are to be ranked, is that FCs are defined in terms of the facts that they hold fixed and allow to vary. They are thus only defined with respect to a starting set of facts (out of which the FC chooses which to hold fixed and which to vary), which I will call the context, Z. The most obvious set of facts to start from is the facts of the actual world now. However, we could also start with the facts at other times, if we want to assess what will be feasible at a future time or was feasible at a past time, or with other possible sets of facts, if we want to assess what counterfactually would be feasible in other possible worlds.[11] An FC holds fixed certain of the facts of Z beyond the time of Z and allows others to change.

The position of an FC on the FC-scale, then, will be a matter of how realistic from Z is a state of affairs in which the facts it varies (f_v) do not hold and the facts it holds fixed (f_f) do.[12] That is, it will be a matter of how realistic it is in Z (when the facts of Z obtain at t_z) that a state of affairs where f_v hold and f_f do not will obtain after t_z. Of course, if we

[10] The qualifier 'almost' is important here. Some facts logically or conceptually imply other facts. This means, for certain pairs of facts, it may not be possible to consider what would be feasible allowing one to change but not the other. For example, if $\sim f_1$ logically implies $\sim f_2$, then it will not be possible to allow f_1 to vary but not f_2.

[11] Strictly speaking the framework disallows such counterfactual comparisons, since FCs are defined by the set of facts (out of those that hold on some given Z) they allow to vary and so are defined only given a Z. This makes feasibility comparisons (which require knowledge of which FCs make which outcomes feasible) across different Z impossible. However, I think that we could make such comparisons while accepting that specific FCs are relative to a choice of Z by recognising that specific FCs on different Z may be of a *type*. We can thus make comparisons across different Z by comparing feasibility on FCs of the same type.

[12] Allowing a fact F to vary or to change means allowing it to hold or not hold. This makes matters slightly more complicated since we cannot just say that, for example, we allow human motivations to vary. The simple fact that motivations are exactly as they are will be relatively easily changed, since it does not take much to make that fact not hold (just change them slightly). The fact that motivations are within range *r* will be somewhat more unchangeable and will become more and more unchangeable as *r* is expanded. This is how we can model the fact that how demanding an FC is will depend not only on whether it allows, say, motivations to change, but also how far it allows them to change. I will talk loosely below of allowing things to vary like motivations or the laws of physics.

consider different times after t_z we will get different results. Generally, as we leave more time after t_z it will become more realistic for a state of affairs significantly different from Z to obtain. If it takes a longer time before some state of affairs becomes realistic, then that state of affairs is thereby less realistic from the current time. If it will be very realistic for some state of affairs A to obtain in 100 years from now but not before, while it is very realistic for another state of affairs B to obtain in only 5 years from now, then B seems to be more realistic than A from the point of view of now. Thus, how realistic a state of affairs is from Z will be a function of how realistic it is at different times after t_z and how long after t_z those times are.

How realistic a state of affairs is from Z comes down to something like the *plausibility* of different possible worlds (the plausibility of the different possible worlds in which the state of affairs obtains at various times after t_z). There are many different possible worlds in which all of the facts of Z obtain at t_z and thus which share the same history up to t_z. These worlds give us the different possible futures of Z. If Z is the facts of the actual world now, then we will not know which of these possible worlds is the actual one, because we do not know what our future is. However, we do have an idea of which of these possible worlds are more plausible than others. A world in which the facts are as they are in the actual world up to the current time, but then shortly after the sun suddenly disappears, for example, seems not to be a very plausible world. It is obviously a complex theoretical matter what these judgements of plausibility are based on and when and to what extent we are warranted in making them. I will just rely, though, on the fact that we are generally able to make such judgements (at least roughly). We would be able, albeit only intuitively, to rank different possible worlds in terms of plausibility.[13] The position of an

[13] It might be thought that there are incommensurabilities in terms of the plausibility of different possible worlds. If this is the case, then my account would have to be more complicated. There would be certain feasibility comparisons that could not be made. However, I cannot think it would pose a problem for the account generally. There are some possible worlds that we can quite

FC on the FC-scale, then, will be some function of how plausible are the most plausible possible worlds at which the facts of Z hold up to t_z and f_v do not hold and f_f do hold at different times after t_z as well as how much after t_z these times are.

There seem to be a few different factors that will be relevant to determining the answer to this question for an FC. Firstly, how hard to change or likely to change the most deeply ingrained fact in f_v is seems important. We can get an intuitive idea for different facts how deeply ingrained they are.[14] The laws of physics, for example, seem to be very deep, unchangeable facts about the world, while people's preferences among ice-cream flavours are quite easy to change. The hardest fact to change that an FC varies is in a sense how demanding that FC requires us to be, it is the most unreasonable thing that it lets vary. Secondly, the number of facts that an FC allows to change seems important. An FC that allows a large number of facts to change is more demanding or more 'utopian', since generally it is more difficult to change a large number of things together.[15] Further, any interaction effects between different facts (that is, effects where changing or holding fixed one fact makes other facts harder or easier to change) will also be relevant. I do not know, though, how these things should be balanced to determine the realisticness of FCs. Thus, without further work, we can only have a rough intuitive idea of where different FCs come on the FC-scale, how realistic or demanding they are. We can, however,

---

clearly say are more plausible than certain others. The account that I will give of 'feasibility' would still allow us to make feasibility comparisons in many cases. If there are incommensurabilities in the plausibility of different possible worlds, then it seems perfectly reasonable to suppose that there are incommensurabilities in feasibility as well.

[14] The question how exactly we determine which facts are more deeply ingrained than others raises some difficult metaphysical questions. Making judgements about feasibility, though, seems to rely on the intuitive notions that we often have that some facts are harder to change or more fixed than others.

[15] Of course there is a problem here about the individuation of facts. There is a very large, or even infinite, number of facts in the world, since there are so many different ways in which the facts can be cut up. We can get around this by saying that if an FC allows to vary any pair of facts *a* and *b* where *a* (logically or conceptually) implies *b*, then the number of facts an FC allows to vary can only count *one* of them. This avoids counting both the disjunctive fact $f_1 \vee f_2 \vee \ldots \vee f_n$ and all of the facts $f_1, f_2, \ldots, f_n$. It also avoids counting separately facts like the fact that human motivations fall within range r_1 and the fact that they fall within wider range r_2.

understand what it means for one FC to be higher or lower on the scale than another. In many cases it will be clear which of two FCs is more realistic, or lower down the FC-scale. For most FCs, I think, it will also be obvious *roughly* where it comes on the scale, that is, whether it is high or low or middling.

Binary or scalar?

Certain philosophers have attempted to find a binary definition of feasibility, a definition such that any outcome can be said to be either feasible or not feasible, giving specific necessary and sufficient conditions for feasibility (for example, Mark Jensen's claim that logical consistency, non-violation of the laws of nature, fixed history of the world and 'natural human ability' are together necessary and sufficient for 'practical possibility', which I take to be an equivalent concept).[16] What all of these theorists miss is that we cannot straightforwardly ask 'is O feasible?' with no further clarification. As we saw above, there is no straightforward answer to the question whether a system of PR is feasible in the UK, unless we know more precisely what we mean by that question. The most sensible answer would be 'it depends what you mean by "feasible"', i.e., 'it depends on which FC'. There are many different possible sharpenings, no single one of which is obviously privileged. The question, then, just as such, does not generally have a determinate answer. We could think about feasibility in terms of a supervaluational structure, meaning that if a proposal is 'superfeasible' (feasible on all FCs) then we can say straightforwardly that it is feasible *tout court* and if it is 'superunfeasible' (unfeasible

[16] E.g. Cowen (2007); Brennan and Southwood (2007); Buchanan (2004, 61); Hawthorn (1991, 158); Jensen (2009) and Räikkä (1998). Gilabert and Lawford-Smith (2012) and Lawford-Smith (2013) distinguish between binary and scalar concepts of feasibility, and give definitions for both. Certain other philosophers have recognised that there are more than one possible way to understand what is meant by 'feasible', that is, different sharpenings of 'feasible', but have given a handful of binary definitions rather than a multiplicity, such as Miller (2008), Brighouse (2004, 27-8) and Elster (1985, 101). It is not any more obvious that there are a handful of privileged binary definitions than that there is just one, since by changing a few of the facts held fixed by an FC we get a slightly different sharpening and it is not obvious why any small handful of these are privileged over the other possibilities.

on all FCs) then it is unfeasible *tout court*. Thus, the question may have a determinate answer if the proposal is either superfeasible or superunfeasible, but most of the time this will not be the case. In order to get a determinate answer to the question, an FC (or range of FCs[17]) must be specified or understood. What is meant by an utterance of 'It is feasible for X to bring about O' could be given by the context. The speaker may in some cases tacitly or explicitly assume a particular sharpening (or range of sharpenings) of 'feasible'. This may result in speakers talking past each other, but need not if the sharpening assumed can be understood by interlocutors. When we say that instituting a system of PR is unfeasible, we presumably do not mean that it is unfeasible holding fixed only the laws of physics. In other cases, it may be left unclear which sharpening is meant, in which case it is indeterminate what the truth conditions of the utterance are. In any case, there is no single sharpening that must be meant by such an utterance. Any FC on the FC-scale is a *possible* sharpening of 'feasible' and, though they are not all things that we standardly *do* mean by that term, as I noted above there is at least a variety of different FCs that we do ordinarily tacitly assume. For this reason attempts to give single binary definitions of 'feasible' are misguided.

In a certain sense, feasibility is scalar or comparative. Proposals or outcomes can be more or less feasible. The lower (i.e., the more realistic) the lowest FC on which some outcome comes out as feasible, the more feasible it is. On each given FC, though, a binary definition of 'feasibility' can be given. That is, once we choose the range of things that we will take as changeable and as fixed, we can make sense of every outcome being either feasible or not. Thus, we can give a very general definition of feasibility as a whole as a scalar concept:

[17] If a range of FCs is specified, then the proposal is feasible if it is superfeasible across that range and unfeasible if it is superunfeasible across that range. If it is neither, then its feasibility remains indeterminate for the specified range.

General Def. An outcome O_1 is more feasible in *Z* than another O_2 in Z_1 iff O_1 is (binary) feasible in *Z* on a lower FC than the lowest FC on which O_2 is (binary) feasible in Z_1.

The variable *Z* here represents the context, the time and possible world from which the FCs are defined, as explained above. Thus, there is a scalar concept of feasibility as such, with no lines that can be drawn to separate the feasible from the unfeasible, but this scalar concept is defined in terms of a binary concept of feasibility on a given FC. The only judgements of feasibility we can make without specifying a sharpening, then, are comparative ones. Below I will attempt to define feasibility given an FC.

Gilabert and Lawford-Smith argue that the role of their binary concept of feasibility is to rule out certain proposals as unfeasible (whereas the scalar concept ranks the remaining, non-ruled-out proposals in terms of comparative feasibility).[18] They claim that there are two sorts of constraints on feasibility, which they call 'hard constraints' and 'soft constraints'. The former are things like logical, nomological and biological constraints, while soft constraints include things like economic, institutional and cultural constraints. The former determine binary feasibility, while the latter determine scalar feasibility. However, any such division must be somewhat arbitrary. As I have suggested, there are harder and softer constraints as we go up and down the FC-scale, but there is no one point along the scale that separates the feasible from the unfeasible. Of course, the point I have made so far is just that there is no privileged binary sharpening of feasibility that gives a general account of the concept. Gilabert and Lawford-Smith's point may be instead that there is one specific FC that is the only one that is relevant to ruling out political proposals. The question of which sharpening(s) of feasibility (i.e., which FC(s)) constrain moral theory will be the topic of my later chapters. I do not think Gilabert and

[18] Gilabert and Lawford-Smith (2007) 813

Lawford-Smith are right that there is any such *single* FC, but in later chapters I will return to this question.

Feasibility for feasibility constraints

Above, then, I have given a general definition of feasibility as a scalar concept *in terms of feasibility-on-an-FC*. I now need a definition of feasibility-on-an-FC, a definition of what it is for something to be feasible given a choice of FC. That is, I need a definition-schema that leaves the choice of FC to be filled in. In order to give a precise sharpening of the binary concept of feasibility that is capable of picking out which proposals are feasible and which are not we need to specify an FC. However, we can give a general definition that explains what this binary concept means.

The first question is what feasibility is *of*. It seems clear that feasibility can be of outcomes, or states of affairs. We might think, though, that actions can also be assessed for feasibility. We might wonder whether it is feasible, say, for me to run to Africa. I think actions can certainly be assessed for feasibility but actions can *be* outcomes. That is, for every action φ there is an outcome consisting in X's performance of φ. We also often talk about the feasibility of *things* like political systems or institutions. Similarly, we can understand such talk as being about the feasibility of outcomes in which those things are in place. Thus, for the sake of simplicity, we can bring all the categories that can be assessed for feasibility under the category of outcomes. The left-hand-side of my definition, then, will be the schema:

(A) O is feasible (for X) in Z on *f*.

where O is an outcome and *f* is some FC. Thus, when assessing this sort of feasibility of some outcome, we must decide on what FC we are considering its feasibility, and then for whom we are considering its feasibility (or whether we are considering its feasibility in

general, not for any particular agent) and finally in what time and possible world. Gilabert and Lawford-Smith also offer a schema including the context as a variable.[19] However, only part of what they wanted to capture is captured by the Z in my schema, part is captured instead by the FC. A high FC can represent a long timescale; if we are concerned with long-term feasibility this is probably because we want to allow more time in which more things might change which might make certain proposals feasible that are not in the short-term, when fewer things can change.[20]

Now G.A. Cohen suggested that there are two elements to feasibility: accessibility and stability.[21] When we ask about the feasibility of a proposal we may want to know whether there is a way that we can get to it from here or whether when we get there it will be stable or both. Sometimes 'feasibility' is used simply to mean accessibility but in other uses it requires *both* accessibility *and* stability.[22] I will focus on the use that requires both, but the account I will give of accessibility should serve on its own as an account of the other use. Whether participatory democracy is accessible is a matter of whether there are paths available to us that will lead to participatory democracy, or whether there are obstacles in our way. Whether it would be stable is a quite different matter; it is the question whether, if the obstacles were removed and we could get there, it could be

[19] Gilabert and Lawford-Smith (2012) 812. Their original schema was: 'It is feasible for X to φ to bring about O in Z'.

[20] Jensen's distinction between synchronic ability (to bring things about now) and indirect diachronic ability (to bring things about later provided one brings something else about first) can in this way be modelled using FCs. Things that are only feasible on higher FCs will be things that we can only bring about diachronically, once we change a number of other things. (Jensen (2009), 173-6)

[21] Cohen (2009) 56-7. This distinction is very similar to Norman Daniels's (2014) distinction between 'achievability' and 'sustainability', which I will discuss in chapter 2 and Erik Olin Wright's (2006) distinction between 'viability' and 'achievability' (97-9).

[22] For this reason David Wiens (forthcoming) argues that stability is not a necessary condition for feasibility (3, n. 2). It seems clear that there is a use of 'feasibility' for which stability *is* a necessary condition, as in when we say that communism is not feasible because human nature will lead it to collapse. Gilabert and Lawford-Smith argue that getting to some outcome, if it cannot be maintained, does not really look like 'getting there' at all (2012, 813). Wiens is right, though, that this is not always how 'feasibility' is used, as in when we say that it is feasible to balance a spinning-top on its point, despite the instability of that position.

successfully sustained. Note that these two things are quite separate. On some FC, an outcome may be accessible but not capable of being stable, or capable of being stable but not accessible. It seems clear that the former may be the case but the latter may be more questionable. If an outcome cannot obtain given *f*, then how can it obtain stably given *f*? The only way we can talk about an outcome being stable but not accessible on some FC is in counterfactual terms. In some cases, this may make little sense. If, say, one of the reasons for an outcome's being inaccessible on *f* was that the outcome itself was incompatible with something held fixed on *f* (as opposed to it's being arrived at being incompatible with something held fixed on *f)*, then we cannot really imagine a counterfactual world in which the outcome already held in which to ask whether its stability would be compatible with *f*. However, in some cases the facts held constant on an FC may be a constraint on what is accessible without constraining what would be stable. It may be, say, that human nature as it currently stands is a constraint that makes anarchism inaccessible, but had we already achieved anarchism, the same traits of human nature would be perfectly consistent with its stability. There may be difficult conceptual problems raised by such counterfactual comparisons of feasibility, but for now the key point is that accessibility and stability are separate.

Accessibility

To begin with, then, I will attempt to give a schematic definition of accessibility. I will return to stability below. Gilabert and Lawford-Smith give a test for binary accessibility which is not intended as a test for accessibility-on-an-FC, but rather for binary accessibility *tout court*. I will use it, though, to help me get my definition. They suggest the following:

> Test 1/Binary: It is feasible for *X* to ϕ to bring about *O* in *Z* only if *X*'s ϕ-ing to bring about *O* in *Z* is not incompatible with any hard constraint.[23]

As I have said above, the notion of a 'hard constraint' is an arbitrary division among FCs. I am interested in a definition of feasibility given the choice of some FC, so I will replace the notion of a hard constraint with the chosen FC, *f*. I will also make the compatibility with the constraints implied by the chosen FC not only a necessary condition for binary accessibility on that constraint, but also a sufficient condition, since accessibility on a constraint is just everything that is allowed by that constraint, everything that is not incompatible with the things it holds fixed. Thus, I propose this definition for binary accessibility given a choice of FC:

> (ARA) O is accessible for X in *Z* on *f* if and only if *X*'s ϕ-ing to bring about *O* in *Z* is not incompatible with constraint *f*/is possible given constraint *f*

where 'ϕ-ing to bring about *O*' means performing some action φ that will bring about O (or will make things such that an event *e* occurs that will bring about O) and is intended to bring about, or to contribute towards bringing about, O and can reasonably be expected to bring about, or to contribute towards bringing about, O. I mean 'not incompatible with constraint *f*' to be equivalent to 'possible given constraint *f*'. It may help to see what is involved in something being possible given some FC to think of an FC as playing a similar role to an accessibility relation in modal logic. An event is possible given an FC if it occurs in some possible world out of a restricted range selected by the choice of FC and Z.[24] The world from which the accessible worlds must be accessible (call this the home world) is selected by Z (it is likely to be the actual world, but need not be). The accessible worlds

[23] Gilabert and Lawford-Smith (2012) 815. The definition is stated as though it is a definition of 'feasibility' rather than 'accessibility', but they clarify that they are concerned only with the accessibility part of feasibility in their discussion.

[24] Note that *e* must be an event that occurs in one of these possible worlds and brings about O. It may be synchronically or only diachronically possible. That is, it must just occur at some point in one of those possible worlds, it need not be immediate from the time of Z.

are then restricted to those identical to the home world up until the time of Z. Finally, the FC then restricts the accessible worlds to those in which, after that time, all facts remain fixed except for those that the FC allows to vary. If an outcome is brought about (directly or indirectly) by X in some possible world out of this restricted range, then it is accessible for X given this FC (and Z). What this means, in less abstract terms, is that when we choose a range of facts to hold fixed, say the deepest facts of human nature along with the laws of physics, biology and so on, an outcome is accessible for me if and only if there is some possible world in which those laws and facts of human nature hold (and which is identical to the actual world up to now) in which I bring about the outcome in question.

The above definition can then be expanded to give a definition of binary *feasibility* on a given FC:

> (ARF) O is feasible for X in Z on *f* if and only if X's ϕ-ing to bring about *O stably* in *Z* is not incompatible with constraint *f*/is possible given constraint *f*.

This, of course, leaves 'stability' to be defined, which will be done below.

Now, feasibility judgements are not about what is *probable*. Something might be feasible (even *very* feasible, that is, feasible on a low FC) but highly improbable. For example, it is presumably fairly feasible for the government to introduce a law banning oranges, but highly improbable, I think.[25] However, if one thinks that there is metaphysical indeterminacy in the world, then it could be necessary to add an element of probability *given the best action* into our definition of feasibility. That is, if indeterminacy is metaphysical then there is an extra dimension of variance to what is accounted for in my definitions above. The best action an agent could perform given some FC to bring about O will not be one that *will* bring about O, but one that will give O a certain probability. Thus, the probability of an outcome *given the best action for that outcome* would also be able

[25] Cf. Wiens (forthcoming, 11, n. 11)

to affect the feasibility of that outcome. Nevertheless, it would still only be the probability of an outcome *given that action* that would be relevant to feasibility; whether an outcome is *overall* probable would not affect its feasibility. If we were to allow metaphysical indeterminacy, we would need to give a scalar definition of feasibility on a given FC and a general scalar definition in terms of that. This could be done in keeping with my broad framework.[26] For the sake of my definitions above, though, (if only for simplicity) I will assume that all such indeterminacy (of whether a given action will bring about a given outcome) is epistemological. That is, I assume away metaphysical indeterminacy.[27]

Now, it might of course seem that probability enters into my definition above in the requirement that for O to be feasible for agent X, it must be possible for X to perform an action that not only *will* bring about O, but also *can be reasonably expected* to bring it

---

[26] We could give a scalar definition of accessibility *on a given FC* along the following lines:

(SARA) O_1 is more accessible for *X* in *Z* on *f* than O_2 is for X_1 in Z_1 on *f* iff
$P(O_1|\varphi) > P(O_2|\psi)$.

where φ is the best action for O_1 for *X* in *Z* (i.e., the one out of those available to *X* in *Z* for which $P(O_1|\varphi)$ is the greatest) that is not incompatible with constraint *f* (the constraint on which we are assessing comparative accessibility) and where ψ is the best action for O_2 for X_1 in Z_1 that is not incompatible with *f*. If we were to take such a probabilistic approach to feasibility *on a given FC* then scalar feasibility *overall* would have to be some sort of function of degree of probability on each FC and the level of each FC on which the outcome in question has a degree of feasibility. This could be done by assuming that a cardinal number between 0 and 1 can be given to every FC according to how high it is on the FC-scale (1 is high, 0 is low) and multiplying the degree of feasibility on each FC by the 'lowness' of that FC. Thus, the overall feasibility of an outcome O for X in Z would be:

$$\sum_{i=1}^{n}\left(1-f_i\right)P\left(O\middle|\varphi_i\right)$$

where f_i is the number between 0 and 1 assigned to FC *i*; φ_i is the best action available for *O* for *X* in *Z* that is not incompatible with FC *i* and *n* is the number of FCs there are. This, then, would be a sum of the results of multiplying the 'lowness' of the FC by the probability of the outcome given the best action for it at *that* FC, for every FC. This, of course, is only one way to get a general definition of feasibility given metaphysical indeterminacy and scalar definitions on FCs; one could give more or less weight to the different elements taken account of.

[27] I do not mean to suggest that, metaphysical indeterminacy aside, there could not be a use for a scalar definition of 'feasibility' *on an FC*. It might be thought that an outcome that is brought about in *more* possible worlds out of those selected by an FC than another is *more* feasible on that FC. (That an outcome comes about in more possible worlds does not mean it is more probable since it is not the case that all possible worlds are equally likely.)

about. It is true that in a sense this brings in an element of probability; it makes feasibility about the *possibility* of performing an action that makes the outcome (subjectively) *probable* (to whatever degree makes it reasonable to expect the outcome). However, feasibility itself is still about possibility (it is about there being *some* possible world that fulfils a certain condition, not about probabilities in the actual world). The inclusion of this requirement deals with the case raised by Brennan and Southwood of a medical ignoramus performing a neurological operation for which they lack the relevant expertise.[28] There is presumably *some* possible world in which, by sheer chance, the medical ignoramus performs all of the right movements to successfully finish the operation. However, with their lack of medical expertise the medical ignoramus could not have reasonably expected the actions they performed to produce the desired outcome. This requirement, though, does not amount to making feasibility a matter of probability conditional on trying, as Brennan and Southwood do.[29] To do so, I think, would be wrong. Consider the case of someone with a pathological phobia of spiders. On an FC that allows to vary all of the agent's motivations, we would want to say that it *is* feasible for the spider-phobic person to hold a spider. My account has this result, as, within the range of worlds selected by this FC, there is one where they perform an action that they can reasonably expect to result in holding the spider. However, the conditional probability view makes feasibility about the counterfactual probability the outcome would have if the agent tried. The only possible world that is relevant to this is the closest one in which the agent tries. In the case of the spider-phobic agent, if they tried, they would most likely not succeed, as their phobia would prevent them. We wanted to allow the agent's motivations to vary, so perhaps we should ask about conditional probability given motivations being different. However, we cannot make sense of this, since to know the probability of success we need to know the facts about the agent's motivations. We

[28] Brennan and Southwood (2007) 8-9
[29] Ibid. 9-10. Gilabert and Lawford-Smith (2012) and Lawford-Smith (2013) do the same.

cannot say what the probability of success conditional on trying would be if the agent's motivations simply were not what they actually are, since they could be anything else.

Now, the definition (ARF) above is a definition of *agent-relative feasibility*. This means that it defines the feasibility of an outcome *for some agent(s)*. We may also, however, want a non-agent-relative definition of feasibility (on a given FC), a criterion for what it would take for an outcome to be feasible *tout court* on some FC (as opposed to feasible *for some X* on a given FC). Lawford-Smith suggests that whether we should assess agent-relative feasibility or non-agent-relative feasibility will depend on the proposal that we are assessing for feasibility. I will base my non-agent-relative definition on Lawford-Smith's definition of binary feasibility, adapting it to be limited to a given FC. The definition she gives is:

> An outcome is feasible iff there exists an agent with an action in her (its) option set within the relevant temporal period that has a positive probability of bringing it about.[30]

To start with accessibility, then, on a given FC, I will adapt her definition thus:

> (NARA) An outcome O is accessible in Z on *f* iff $\exists X \exists \varphi$(X's φ-ing is not incompatible with constraint *f*/is possible given constraint *f*).

where *X* is an agent and φ is an action that will bring about O (or will make things such that an event *e* occurs that will bring about O) and is intended to bring about, or to contribute towards bringing about, O and can reasonably be expected to bring about, or to contribute towards bringing about, O. One might think that a non-agent-relative definition of accessibility ought not to involve reference to agents and actions at all; an action is accessible on some FC just if there is a possible event that would bring it about compatibly with that FC. However, I think that feasibility is about what can be done. If something is possible, but cannot be brought about by any agent, then it is not *feasible*, it

[30] Lawford-Smith (2013) 250

is merely *possible*. This is the distinction between the two concepts, feasibility and possibility; the former requires agency while the latter does not. Feasibility is not just equivalent to possibility. Again, as in my discussion of (ARA) above, this definition can be understood in terms of possible worlds. The existential quantifier quantifies over the restricted set of possible worlds selected by Z and the chosen FC together. The requirement of (NARA) is that there be an action that brings about O in at least one of these possible worlds. I have removed the element of probability from Lawford-Smith's definition since above I assumed away metaphysical indeterminacy and the temporal period restriction since this is captured by the choice of FC.[31] To take one of our examples, then, participatory democracy is accessible on an FC that holds fixed certain deep facts of human nature only if, compatibly with those facts, it is possible for some agent(s) to bring it about. This does not mean that there must be any one agent who can single-handedly bring it about. All that is necessary is that there is an agential route to the outcome; if the outcome is possible but only through non-actions then, though it is possible, it is not feasible.

To include stability in this definition and get a non-agent-relative definition of feasibility all we need to do is add that O must be brought about *stably* (again, the meaning of this will be expanded on below).

Now, of course, the distinction between agent-relative feasibility and non-agent-relative feasibility applies also to the general scalar definition of feasibility that I gave at the start. Thus, we can split the general definition into two definitions, agent-relative:

> (GARF) An outcome O_1 is more feasible for *X* in *Z* than another O_2 is for X_1in Z_1 iff O_1 is (binary) feasible for *X* in *Z* on a lower FC than the lowest FC on which O_2 is (binary) feasible for X_1in Z_1.

[31] Again, as with agent-relative accessibility, there is also a scalar sense of non-agent-relative accessibility (and feasibility).

and non-agent-relative:

> (GNARF) An outcome O is more feasible in *Z* than another O_2 in Z_1 iff O is (binary) feasible in *Z* on a lower FC than the lowest FC on which O_2 is (binary) feasible in Z_1.

Stability

Now, then, I turn to stability, the element of feasibility that I have so far included without further explication. There is limited discussion in the literature of how to define 'stability'. Rawls put some importance in the stability of his conception of justice, and he did attempt to clarify what this entailed.[32] Rawls defines stability for *systems*, whereas what I want is a definition of stability for outcomes, or states of affairs. However, what he says regarding systems will be useful as a point of departure. He says that stability for systems is a matter of the forces in the system that will return the system to *equilibrium*. A system is in equilibrium 'when it has reached a state that persists indefinitely over time so long as no external forces impinge upon it'. An equilibrium is stable 'whenever departures from it ... call into play forces within the system that tend to bring it back to this equilibrium state'.[33] Rawls thus requires that departures from a stable system must *themselves bring about* a return to equilibrium. Thus, when we consider an outcome, the equivalent requirement would be that departures from that outcome tend to *bring about* a return to that state of affairs. I do not see, however, why we must require this. Presumably an outcome would be stable whether departures from it tended to *bring about* returns to equilibrium, or whether departures were simply followed by returns.

Thus, my definition of stability is the following:

[32] Rawls (1999) 398-400
[33] Ibid. 400

(S) An outcome O is stable iff it will be maintained as an equilibrium, with small departures from equilibrium being subsequently corrected.[34]

There is obviously some vagueness here, since the question how frequent, extensive or pervasive departures from an outcome must be before we determine that that outcome is not stable is not given any clear answer. This, though, gives us a binary notion of stability. There is also, however, a scalar notion. Outcomes that are not stable in this demanding binary sense may be more or less stable. Outcomes that are not sustainable indefinitely, but for relatively long periods of time approximate more to stability. An outcome that can only be sustained for, say, a day is less stable than one that can be maintained for long periods (years, decades perhaps), but that will eventually collapse.

Stability is not a modal notion (unlike accessibility and feasibility). Making claims about stability does not involve making claims about other possible worlds. An outcome is stable just if it will be maintained. However, we are often interested not in the stability of existing states of affairs, but in the stability of proposed outcomes. These sorts of claims about stability are counterfactual (modal) ones. What we want to know is whether there is a sufficiently close possible world in which the proposed outcome is implemented and is stable, in other words whether it is capable of being stable. The question whether some outcome is *capable* of being stable is susceptible to numerous different interpretations, in the same way as questions about accessibility are. That is, there are many different FCs on which it could be made precise. We may want to know whether an outcome is capable of being stable given the current political system, balance of power and so on being fixed, or alternatively we may be interested in whether it is capable of being stable given only some of the deeper facts of human nature. Modal facts about stability are also important

[34] A repeating cycle or something that changes within some parameters can be stable. If we define the outcome whose stability we are asking about broadly, such that the outcome can persist through changes within some parameters, then the stability of this outcome is consistent with change.

to claims about *feasibility*. To assess the feasibility of an outcome (when feasibility is taken to require stability) we need to know another specific modal fact about its stability. As I have said above, an outcome O is feasible in Z on *f* if and only if it is accessible *stably* in Z on *f*. For O to be feasible, it must be that the outcome 'stable O' is accessible.[35] Thus, O is feasible in Z on *f* if and only if there is a possible world out of those selected by *f* and Z in which O is brought about and is stable. O must be stable *in the same possible world* in which it is brought about. For example, then, for anarchism to be feasible given certain facts about human nature being held fixed, those facts must both not prevent anarchism from being brought about, and also not prevent it from being maintained whenever it is brought about.

Motivations

This completes my definitions of 'feasibility'. However, before I move on to discuss how this affects moral theory, I will briefly discuss two other important ideas in the literature about how 'feasibility' should be defined. Lawford-Smith claims that the motivations of an agent should not count as a constraint on the feasibility of an outcome for that agent. This is because to do so would to be to make too much infeasible. We should not let an agent off the hook morally because they are not motivated to do the thing in question. Thus, if infeasibility is a defeater for moral duties, then we should not say that something is infeasible for an agent simply because that agent is not motivated to do it.

However, I think we can think about this differently when we do not see feasibility as a binary matter. The best way to think about this question is to allow that a choice of FC can relativize its choice of facts to be allowed to vary to the agent whose ability to bring

[35] Note that stability can be reduced to accessibility. If, instead of asking about the accessibility of outcome O, we ask about the accessibility of outcome 'stable O' we no longer have any need to ask separately about stability. However, ordinarily when we ask about the feasibility of some outcome O we want to know about *both* accessibility and stability, so we do not want to make stability questions separate questions to be asked in terms of accessibility.

about an outcome is being assessed (if we're talking about agent-relative feasibility). In certain cases the facts that a particular FC selects to be held constant may vary as we change the agent whose ability to bring about an outcome we are assessing. Thus, we may want to consider whether some outcome is feasible for some agent on an FC that holds constant the motivations of *others* (and only others, that is, the motivations of the agent should be allowed to vary). The facts that this FC holds constant vary as we change the agent we are considering. On such an FC, we can then say that an outcome O is *non-agent-relative* feasible if there is some agent X for whom it is agent-relative feasible (that is, if there is some agent whose φ-ing to bring about O is compatible with the motivations of agents other than them).

If we allow FCs to hold constant facts that are relativized to agents in this way, we can avoid Lawford-Smith's wholesale ruling out of agents' motivations as constraints on feasibility.[36] There will be one FC on which we can assess the feasibility of an outcome which holds constant the motivations of other agents, but not of the agent in question, as Lawford-Smith wants to do in general. There will also, though, be a lower FC on which certain of the agent's motivations are held constant and not others, and even one on which all of the agent's motivations are held constant. Lawford-Smith notes that there is a continuum between failing to try and motivational pathologies like addictions, phobias and illnesses. It would be preferable if we can allow for this continuum rather than having to draw a fine line through it. On my account we can have a variety of FCs that allow decreasingly pathological lack of motivation to count as constraints on feasibility. Thus, an FC that holds constant addictions and phobias, say, will be higher up the FC-scale than one that holds constant all the agent's motivations. There will presumably be a range of FCs between the two that hold constant fewer and fewer motivations as they become

[36] Wiens (forthcoming) notes that 'some ordinary feasibility claims seem duly attentive to at least *some* motivational constraints' (5).

more and more pathological. Lawford-Smith is right that we let an agent off too easily if we let them off anything that they are not actually motivated to do, but there is no need to include this fact in the *definition* of 'feasibility'. An FC that holds constant even the agent's own motivations is a sort of limiting case. We will rarely perhaps be interested in what is feasible on such a low FC for moral philosophy, but it is still *a possible* understanding of feasibility.

Moral costs

Juha Räikkä argues that a definition of feasibility itself should include 'the necessary moral costs of changeover', that is, the moral costs of getting to the outcome being assessed for feasibility, as a constraint on feasibility.[37] He thinks that an outcome can be made unfeasible by its moral costs. If getting to the outcome is too morally costly, then this may mean that the outcome itself is not feasible. However, I think this is a mistake. Gilabert and Lawford-Smith distinguish between *believed* moral costs and *actual* moral costs.[38] Of course, believed moral costs may be a constraint on feasibility; certain FCs will hold fixed such beliefs and they may impact on what is feasible on these FCs (if people believe that the moral costs of a proposal outweigh its benefits then they may not want to pursue it and so if their cooperation or support is necessary to achieve it, this may prevent it from being feasible on these FCs). Actual moral costs, though, do not factor into a definition of 'feasibility'. The question of whether some outcome is feasible *in conjunction with* acceptable changeover costs is a different feasibility question to whether the outcome is feasible tout court. As I argued in the introduction, the fact that an outcome is feasible does not imply that it is feasible in a desirable way. Räikkä argues that the feasibility of a proposal is the feasibility of the *successful* implementation of that proposal. This is true in the sense that it is only feasible if there is an agent who *can* try to

[37] Räikkä (1998)
[38] Gilabert and Lawford-Smith (2012) 817

bring it about and actually do so (this is a basic sort of success). Further, it must also be possible for it to be implemented successfully in the sense of stably. However, Räikkä understands 'successful' as being moralised, such that a successful implementation of a proposal is not only a stable implementation, but a morally desirable one. I have already argued that the question of whether some proposal is feasible in an all-things-considered desirable way is a different question to whether it is feasible simpliciter. It is true that in one sense of 'successful', an implementation of a proposal would have to be all-things-considered desirable to count as successful. However, to require successful implementation in *this* sense of 'successful' is to conflate the desirability and feasibility questions into one question: a proposal is then only feasible when it is *both* desirable *and* feasible. This seems clearly wrong.

Of course, whether the all-things-considered desirable implementation of a proposal is feasible (that is, the conjunction of the proposal and the realization of whatever other values are considered necessary to make it all-things-considered desirable, including the process of changeover) are still questions we can, and should, ask, but they are distinct questions from whether the proposal is feasible tout court.[39] We thus need to ask feasibility questions not just about *culmination outcomes*, to borrow terminology from Sen, but also about *comprehensive outcomes*.[40] Comprehensive outcomes include the process involved in getting to the culmination outcome as well as all other aspects of the world. If, say, we are interested in the proposal of open borders, then, we need to be interested not just in whether that proposal is feasible, but also in whether it is feasible to bring about a desirable comprehensive outcome involving open borders: that is, one in which whatever is desirable about open borders is not outweighed by moral costs of changeover or by undesirable aspects of the culmination outcome.

[39] Buchanan (2004) calls the combined question of feasibility and desirability 'moral accessibility' (61). Stears (2005) brings out the importance of assessing moral costs of changeover.
[40] Sen (1997) 745. He applies this distinction to politics in (1999) 27.

We should distinguish also, as Cohen does, between a principle or proposal and its implementation.[41] The evaluation of these two things is separate, which Räikkä misses. It may be that some ideal world is morally desirable, but that the only way it could be brought about would be through means that would impose such significant moral costs that they would outweigh the desirability of the ideal world itself. This would imply that the ideal world *is* morally desirable, but its implementation (in a feasible way) is *not*. Further, a proposal may be morally desirable and feasible, yet its implementation not be morally desirable because its implementation *in a morally acceptable way* is not feasible. I do not think, then, that the 'moral costs of changeover' of an outcome (or any other evaluative or normative feature of it) is a part of its feasibility. They do not enter into the definition of the latter term, though they are important in deciding which feasibility questions we should ask.

On my account of feasibility it might seem that moral costs of some proposal *could* count as a constraint on its feasibility if there are moral facts. If there are moral facts, then they could be included as facts held constant on some FC. However, even if we did hold constant moral facts it is not clear that this would change the feasibility of any proposals, since moral facts (if there are such) would not seem to be the sort of facts that could make any outcome defined in non-moral terms *impossible*. They could make things impossible as morally acceptable actions or outcomes (and thus could make impossible an outcome that is defined in terms of moral acceptability). When I say, then, that we should assess the feasibility of implementing outcomes in an all-things-considered way, I do not mean that we should define outcomes in terms of all-things-considered desirability (for example, define an outcome 'the all-things-considered desirable implementation of open borders' and ask whether it is feasible). Rather, we should

[41] Cohen (2009). Gilabert argues similarly that when we *are* deciding what proposals to implement we will have to consider the desirability and feasibility of the *implementations*, not only of the principles (Gilabert, 2011, 61-63).

consider which comprehensive outcomes involving the proposal in question are desirable and assess the feasibility of these non-morally defined outcomes (e.g. determine that open borders would be desirable if combined with *x*, *y* and *z*, and then ask about the feasibility of the comprehensive outcome consisting in open borders and *x*, *y* and *z*).

Chapter 2

I will now turn to the question that will occupy me through the rest of this thesis. That is the question of what facts about feasibility (given how I have argued we should understand that concept) imply for moral theories, in what ways they constrain what morality can require of us, and in what ways we ought to take them into account when doing moral and political theory. The guiding aim of the thesis is to gather some tools to investigate when and how (or whether) feasibility critiques can be successful arguments against some moral (or political) theory.[42] What we want to know, then, is when a fact about the feasibility of some state of affairs implies that morality cannot demand we bring it about or that we should not be interested in theory that recommends it.[43] I will consider the first part of this without giving a proper answer, but in the third chapter I will argue for a view regarding the second part (though this will only give a vague and negative answer).

I argued in the first chapter that there is a multiplicity of different possible sharpenings of the term 'feasibility', no single one of which is obviously privileged (and the facts about feasibility are different on different sharpenings). What this straightforwardly implies is that, in arguing against a theory or proposal, it is not sufficient simply to say that it is unfeasible. What that means is indeterminate, and it seems clear that unfeasibility will not be grounds for rejecting a theory or proposal *whatever* sharpening of the term we adopt. Given what I have argued, a feasibility critique cannot just reject a theory by saying that it is not feasible simpliciter, but one could argue that a theory is not correct or interesting because it is not feasible on some specific sharpening, *f*. The question now

[42] I mean 'political theory' only to refer to normative moral theories about the political realm.
[43] I mean 'demand' in a weak sense that is neutral between a strong *requirement* (what morality says we *must* do) and a demand that we must comply with in order to be morally perfect (what morality says we *ought* to do, even if we are not *required* to do it). There is also an evaluative form of moral and political theory that is not about what we normatively should do or bring about, but rather about what would be good. It seems fairly clear that this evaluative sort of moral theory is not constrained at all by feasibility, so I focus on *normative* moral theory.

arises, then, whether there is a single sharpening, or a number of specific sharpenings, that constrain moral theory or the sort of moral theory that we ought to be interested in, while the others do not. If there is such an FC (*f*), then it would be sufficient to object to a theory just to say that it is not feasible on *f*.

Norman Daniels, in a recent colloquium paper, set out a puzzle about the relation between feasibility facts and moral theory.[44] He focuses on the specific subdomain of morality that has to do with justice, but much of what he has to say is intended to be relevant to morality more generally. He noticed that in some cases the judgement that some proposal is unfeasible does not prevent us from concluding that it is a requirement of justice, whereas in other cases it does. The first type of case he calls an 'A case': 'in these cases ... the infeasibility of instituting a just practice, policy or institution does not lead us to revise what we think is just. Rather, we condemn it for its injustice whether or not we can alter its injustice'. He gives as an example of such a case one where we condemn someone's racist treatment by some institution even though we cannot alter that institution. The second type of case he calls 'B cases': here 'the infeasibility of instituting a more just practice, policy or institution leads us to revise what we think a just outcome requires'. For example, he says, 'we might think that highly altruistic people would not engage in a certain practice, but we recognise that people in general are not at all that altruistic and so we may reasonably conclude that justice cannot require people to avoid that practice'. The puzzle, then, for Daniels, is to explain why in some cases unfeasibility seems to constrain what justice (and morality) can require, while in others it does not. Daniels is thus attempting to answer my question, when or whether facts about feasibility constrain the demands of morality and when they do not. What is needed is thus an account that can explain which facts about feasibility (or which facts in which circumstances) are relevant to constraining moral theory. It could be that the division

[44] Daniels (2014)

among feasibility facts that are relevant and those that are not has something to do with the different FCs on which the facts hold, but Daniels has a different answer. Daniels's answer, if correct, would also seem likely to have implications for which sorts of theory we should be interested in.

Daniels's solution

Daniels's solution to the problem is to distinguish two different sorts of feasibility, one of which he thinks constrains what morality can require of us and one of which does not. In other words, he thinks that the sort of feasibility that is in play in A cases is different to the sort of feasibility that is in play in B cases. The two sorts of feasibility he distinguishes he calls 'sustainability' and 'achievability'. 'Feasibility as sustainability' is about whether some proposal can be sustained by humans as they are. This is contrasted to 'feasibility as achievability', which is about whether some proposal can be achieved now. He argues that sustainability is a constraint on a requirement of justice, while achievability is not.[45] Thus he thinks that A cases (such as the case of the racist institutions that we cannot remove) are ones where the proposal is sustainable though it is not achievable, while B cases (such as proposals made unfeasible by insufficient human altruism) are ones where the proposal is not sustainable.

This distinction seems to correspond closely to Cohen's distinction between accessibility and stability and it is an important one.[46] However, the use Daniels puts it to is different and the idea that it can account for the difference between feasibility facts that constrain

[45] Gilabert (2008) also distinguishes between accessibility and sustainability and argues that the two play different roles (413-4). He does not say, like Daniels, that one constrains moral requirements and the other does not, but argues that they constrain different sorts of theory. He says that the design of institutional schemes implementing the fundamental principles should be constrained by sustainability while the design of processes of reform leading to the realisation of these schemes should be constrained by accessibility. He is probably right that the latter need only be constrained by accessibility; sustainability is irrelevant for one-off actions or processes. However, I do not see why the former should be constrained only by sustainability and not accessibility, for reasons similar to those I will give against Daniels's proposal.

[46] Cohen (2009) 56-7

moral requirements and those that do not seems mistaken. For one thing, Daniels's distinction misses the fact that there are various possible sharpenings of 'feasibility' (i.e., different FCs). This variety holds for both sustainability and achievability as well as for feasibility itself. What is sustainable on certain low FCs might be unsustainable on other higher FCs.[47] For instance, some proposals may be unsustainable on some FCs because human motivations as they currently are tend to tempt people away from the proposal, but they could be sustainable on other higher FCs where these motivations are not held fixed. Thus, Daniels's distinction does not help answer the question what sort of strength of feasibility (i.e., what FC) ought to constrain moral requirements. He presumably will not want to say that one cannot be required to do anything that is unsustainable on very low FCs (that hold fixed a very large range of facts). Equally, presumably the sustainability requirement is stronger than just requiring sustainability on *some* FC, since the highest FCs allow a very large range of facts to change and thus make all sorts of proposals sustainable; it seems likely that there is some FC significantly below the top of the FC-scale which the observance of moral requirements must be feasible below. Achievability, too, can be interpreted in numerous ways and it seems unlikely that even very high-FC unachievability cannot be a constraint on moral requirements.

In any case, the distinction between sustainability and achievability does not seem to correspond at all to the distinction between feasibility considerations that should constrain moral theory and those that should not. Sustainability is not always a constraint on moral theories and achievability sometimes is. Riz Mokal, in the discussion of Daniels's paper, questioned whether we need sustainability at all as a constraint on the requirements of justice. If we suppose that it is achievable for us to abolish slavery, but

[47] I take the notion of sustainability to be essentially equivalent to the notion of stability that I discussed above. As I explained, when we are interested in the stability or sustainability of *proposals*, we are interested in whether these proposals are *capable* of being stable. This question requires an FC to be specified in order to have a determinate answer.

there is a tendency over the long term to revert back to slavery-supporting institutions, that is, that anti-slavery institutions are not sustainable over the long term, it does not seem like this latter fact gives us reason to think that justice cannot require us to abolish slavery. It seems, rather, that justice would require us to abolish slavery and then, if slavery re-emerges, so be it, or if when it does it is again achievable to abolish it, we would be required to do so.

On the flipside of this, it does not seem that achievability cannot be a constraint on what we are required to do. It seems like, if there are proposals that are unachievable, but that would be sustainable if they were achieved, their very unachievability *can* make it the case that we could not be required by justice to bring about those proposals, especially if the obstacles to the proposals' achievability were strong. Consider, for instance, the proposal that we should find another habitable planet in the universe and inhabit it. This, if achieved, would be sustainable, since, by hypothesis, the other planet would be habitable and so we could suppose, once it is inhabited, it could remain so. However, the prospects for finding another habitable planet given the current state of science are slim (and there may not even be one). Thus, unless we assume a fairly expansive FC that allows our scientific capabilities to expand a great deal, it seems reasonable to say that this proposal is unachievable. It seems that *because* this proposal is so unachievable, and despite the fact that it *is* (would be) sustainable, it cannot be (at least one sort of) requirement of morality. (There are different sorts of moral requirements, which may well have different relations to facts about achievability. For instance, it may well be the case that the unachievability of the above case does not prevent us from having a *pro tanto* moral requirement to carry it out. However, it seems clear that there at least *exists* a sort of moral requirement that is constrained by some sharpening of achievability. I cannot be *all-things-considered* required to find and inhabit another habitable planet tomorrow. Furthermore, it does not seem like sustainability and achievability constrain

fundamentally different types of moral requirements. It does not seem like, for instance, sustainability is a constraint on pro tanto moral requirements, while achievability is only a constraint on all-things-considered requirements. *If* the unsustainability of an inhuman level of altruism *prevents* it from being pro tanto morally required, then I see no reason not to think that the unachievability of finding another habitable planet also prevents that from being pro tanto required). It is not the case that, as Daniels seems to think, something's being humanly unachievable implies it's not being humanly sustainable.

Now, of course, the term 'requirements of justice' that Daniels uses is ambiguous. In one sense it means the moral requirements that justice puts on us. I just argued that in this sense, unachievability can be a constraint on the requirements of morality (or at least on certain of the requirements of morality). In another sense, though, it means what is required for some state of affairs to qualify as just. In this sense, Daniels could argue that achievability does not constrain what justice *is* (what is required for some state of affairs to fall under the concept 'just'). However, if this is what Daniels is claiming, it is unclear why we should think that sustainability *does* constrain what justice *is*. If what justice fundamentally consists in does not need to be something that is achievable, why suppose it must be sustainable?

Part of the appeal to Daniels of using this distinction to account for the different ways in which feasibility facts can affect moral requirements seems to be the idea that the deeper truths of human nature affect what is sustainable or not, while achievability is to do with how things are at the moment (e.g., the current balance of power, people's current motivations and so on). This, however, I think is wrong. The distinction between the more intractable facts of human nature and those facts that are more easily changed (less permanent or unavoidable) does not correspond to the distinction between achievability and sustainability. Some proposals are unachievable because of the current state of play

(so to speak), because of short term, changeable facts, but others are made unachievable by deeper, more intractable facts. On the other hand, while proposals can be made unsustainable by deep facts of human nature, they can also be made unsustainable *now* by shallower, more temporary facts.

Daniels goes on to argue that his distinction gives a plausible account of '"ought" implies "can"' (OIC). He seems to argue that the sort of 'can' implied by 'ought' is sustainability. It is true that if something is not sustainable then this may be good enough grounds to say that we cannot do it. However, it also seems perfectly natural to say sometimes that because something is unachievable, we cannot do it. It is unclear why sustainability should be considered a constraint on 'ought' and not achievability.

His view seems to be that even if some proposal is unachievable for some individual at the moment, they can still be required to act in accordance with this proposal so long as it is *in general* sustainable for humans. This leads him to the strange view that people can be held responsible for failing to do things that they could not have done if they are things that people in general are capable of doing. Thus he says that 'we may charge (and possibly punish) psychopaths who lack certain capabilities to do what justice requires even though individually they may not be able to because enough other people can behave in the required ways to support the view that such behaviour is sustainably feasible for humans'. This, to me, just seems implausible. If OIC is going to be of any importance at all, it needs to constrain what people ought to do according to what *they* cannot do, not according to general human capacities.[48] When evaluating a person's action, it is irrelevant what other people are able to do.

[48] Of course, as a matter of practicality when designing laws, it may sometimes be best to punish people for failing to do things that they should not have done when people in general can do that thing, because of the impossibility of making laws fine-grained enough to track people's abilities perfectly. However, this cannot be the case for the general meta-ethical principle OIC.

The nature of the feasibility constraint on moral theory

Thus, it seems that it is not the distinction between achievability and sustainability that explains when feasibility does constrain the requirements of morality and when it does not. It also appears that this distinction does not give us the answer to the question when and how moral or political philosophy ought to take feasibility considerations into account. If Daniels was right and the requirements of morality were constrained by sustainability but not by accessibility, it might seem (though it does not follow immediately) that in order to be useful or worthwhile a political or moral theory must meet some sustainability requirement, though it need not meet any accessibility requirement. My arguments above seem to suggest that this is not the case. A theory can be made useless (at least for certain purposes) by accessibility considerations: a theory recommending that we find and inhabit another habitable planet, for example, would seem to be useless at least for many purposes of theory. On the other hand, the unsustainability of an outcome recommended by a theory is not necessarily enough to make that theory useless. Thus, we will need some other sort of argument to establish the feasibility-related limits on useful theory.

I suspect that the solution to Daniels's puzzle may in fact relate to the variety of different sharpenings of 'feasibility'. The difference between those proposals whose infeasibility appears to lead us to reject them as moral requirements and those for whom infeasibility appears not to be a constraint may be explained by the infeasibility facts holding on different FCs. Feasibility facts given different FCs are likely to have different implications for moral theory.

There are different sorts of moral requirements and they may not all be constrained by feasibility (it may be possible for us to be under certain sorts of moral requirements entirely regardless of whether or not it is feasible to fulfil those requirements). However,

it seems hard to deny that there are at least certain cases in which the unfeasibility of some proposal (or its degree of unfeasibility in the scalar sense) is enough to show that a certain sort of 'ought'-claim demanding it cannot be true. For example, in a case where I am on my own having to decide which of two sets of people to save, and where, holding fixed the laws of nature and human strength and speed, it is not feasible to save both sets, morality surely cannot say that *all-things-considered* I ought to save both sets (and this seems to be due to the feasibility fact). Thus, though an '"ought" implies feasibility' (OIF) principle may or may not hold in general, there are at least certain types of case and types of 'ought' for which some sort of feasibility facts are constraints on the truth of 'ought'-claims.[49]

However, given what I have already argued, any sort of OIF principle that does hold (however limitedly) is not at all straightforward. Though 'ought' does seem (sometimes) to imply certain feasibility facts, it will not be enough to say simply that certain sorts of 'ought' imply feasibility *tout court*. I have claimed that there is a multiplicity of different available sharpenings of 'feasibility', any of which could conceivably be taken as a precisification of the term 'feasible' in an OIF principle. It presumably will not be the case, for any type of case and type of 'ought', that 'ought' implies 'feasible' on *all* FCs (i.e., all possible sharpenings). If the former were the case for some sort of 'ought', it would mean that a proposal must be feasible *however* we sharpen 'feasible' (i.e., *whatever* range of facts we hold fixed) in order for it to be possible that we ought to carry it out. This is obviously ridiculous, since one possible FC is one on which we hold fixed all of the facts of the world. On this FC *nothing* is feasible apart from the status quo. It presumably will also not capture the constraint that feasibility imposes on moral theory (in those cases where it does impose a constraint) to say that 'ought' implies 'feasible' on just *some* FC. This would mean that in order for an 'ought'-claim (of the relevant sort) to be true, there must

[49] Such a principle is discussed by Brennan and Southwood (2007).

just be *some* way of sharpening 'feasible' on which it is feasible. This is no doubt true, but it is far too weak. One of the FCs available allows all of the facts of the world to change (arguably excluding the laws of logic). Almost any proposal will be feasible on this FC. There seem to be, at least sometimes, feasibility constraints on moral theory that are much stronger than this.

Thus, even if (or even when) feasibility does impose a constraint on morality, this cannot be captured by the simple OIF principle. Presumably, 'ought' will imply certain sharpenings of 'feasible' and not others. Different sorts of 'ought' (or different sorts of case) may imply different sharpenings of 'feasible'. It could be that for each 'ought'-claim there is a determinate FC that constrains its truth. Alternatively, it could be that each 'ought'-claim does not imply a unique FC, that which FC(s) constrain the truth of any given 'ought'-claim depends in some way on the context of utterance or on the context to which the claim applies. Or it could even be that all 'ought'-claims imply the same sharpening of 'feasible'. It still remains undecided *which* sharpenings of 'feasibility' constrain which 'ought'-claims when. Similarly, if there is a feasibility constraint on the usefulness or value of political or moral theory, it is not obvious on which FC the recommendations of a theory must be feasible in order for that theory to be worthwhile, or on which FCs they must be feasible for theories with different aims to be worthwhile.

I should note, as an aside, that appeal to an '"ought" implies "can"' principle will not help here to determine which sharpening(s) of 'feasible' constrain moral theory. This is because 'can' appears to be imprecise in a way analogous to 'feasible'. Whatever account we give of the sense (or the various senses) of the 'can' of capability or practicability (as opposed to other senses of 'can', such as the 'can' of moral entitlement),[50] just as with 'feasibility', there seem to be various different ways of making its meaning precise. Just as

[50] There have been debates over how to understand the meaning of this sense of 'can'. See, for example, Frankena (1950), Russell (1910, 34) and Austin (1979).

there is not a single set of outcomes that are binary feasible since there are various degrees of expansiveness that we could opt for in our definition of 'feasible', there is not a single set of actions of which it is right to say that we can do them, since we can choose to be more or less expansive with our definition of 'can'. Consider: I am sitting on my bed and my bag is two metres away. Can I touch my bag? Holding fixed the fact that I am on my bed, I cannot touch it. On the other hand, allowing that fact to change, I can. It seems perfectly reasonable to assert either that I can touch my bag, or that I cannot, obviously on different sharpenings of 'can'. In order to get a determinate answer to whether I can or cannot we would need to specify given what facts we want to know whether I can touch it. Or consider: 'The L party cannot enact a law banning alcohol', where the L party wants to enact such a law but does not have a parliamentary majority (which, suppose, is necessary to enact such a law). Holding fixed the composition of parliament and the policy preferences of the parties, the L party cannot enact a law banning alcohol. However, allowing these things to vary, it seems likely that it *can* enact such a law. Feasibility and ability are in fact very similar concepts and thus subject to exactly the same imprecision.[51]

A multiplicity of possible principles

It is clear, then, that the truth of certain 'ought'-claims, at least in certain contexts, is constrained by the feasibility of what they demand on *some* sharpening of 'feasibility'. However it is not clear *which* sharpening of 'feasibility' is implied by *which* 'ought'-claims in *which* contexts. There are at least two different questions though about the constraints that feasibility puts on moral theory, as I have already hinted. One question is the one just mentioned: which feasibility facts (on which FCs) constrain the *truth* of which 'ought'-

[51] '*A* can do *X*' requires 'it is feasible for *A* to bring about a state of affairs consisting in *A* doing *X*', while 'it is feasible for *A* to bring about *X*' is equivalent to '*A* can perform an action that brings about *X*'.

claims or constrain which 'ought'-claims actually hold for us.[52] A certain 'ought'-claim in a certain context may depend for its truth on some feasibility fact holding on some FC. Another question is what feasibility requirements a theory must meet in order to be useful, worthwhile or important. Now of course if the aim that political or moral theory should have is simply to identify the principles and 'ought'-claims that are true, then the answer to this question will be the same as the answer to the first. In the rest of this thesis, though, I will focus on the question of which FCs political and moral philosophers ought to take as constraints on their theory *insofar as they aim to provide moral guidance for action*. This is a normative methodological question. I will set aside the more fundamental question about the truth of 'ought'-claims. I do not wish to deny that there may be a 'truth-seeking' role of moral philosophy, which seeks to identify what the demands of morality are, regardless of whether or not doing so provides any sort of action guidance.[53] I merely set this aim for moral philosophy aside. I think it will be interesting and useful to know which sorts of feasibility constraint moral and political philosophers should take as constraints on their theory when they aim to provide action guidance. Which forms of theory are useful for action-guidance could be independent of which provide true 'ought'-claims. It may be that there are certain sorts of moral principles that do hold, but whose identification could be of no help in giving real people guidance as to what to do in the circumstances in which they actually find themselves. On the flipside, it could be that there are certain sorts of 'ought'-claims we could make that are not *true* (other than counterfactually), but are useful for determining action-guiding 'ought'-claims that are true. The question which FCs constrain the theories we should be interested in is at least potentially independent of the question which FCs can be constraints on the truth of 'ought'-claims. In the third chapter I will argue for a view

[52] I will talk simply about the *truth* of 'ought'-claims. This is only to save space; I do not mean to assume that 'ought'-claims are truth-apt. If they are not, the question will be about the holding of 'ought'-claims.

[53] Cohen (2008, 268) and Estlund (2011) argue in favour of this as an aim of moral philosophy.

relating to the former question. I shall not attempt to provide any sort of complete answer to this question, but will just provide some tools for thinking about it and will argue against a view that claims that the interesting FCs are limited to low or realistic ones.

Now, different FCs rule out different sets of outcomes as unfeasible. If we *assume* a given FC as a constraint on a certain type of moral theory (that is, we assume, counterfactually, that no correct 'ought'-claims of the relevant type can demand that we bring about outcomes that it is unfeasible, on that FC, for us to bring about), it will rule out a certain set of principles, because their observance is unfeasible on that FC. There is, I suppose, for each FC a set of principles that is the correct set of moral principles of some type *in so far* as that FC (and no other) is assumed to constrain moral theory of that type. These are those principles that *would* be true if the FC in question was a constraint on a certain sort of 'ought'. This set of moral principles will represent the best we can do morally given that FC. Since different FCs rule out different sets of outcomes as unfeasible, it is likely that there will be different sets of moral principles (and 'ought'-claims) that will come out as correct when different FCs are assumed as constraints on moral theory, that is, *in so far as different FCs are taken as hard constraints*. I will call these principles that would be correct if an FC, *f*, were a constraint on a certain sort of 'ought'-claim, 'principles-on-*f*'. Since FCs are defined for a particular Z, these will in fact be principles-on-*f*-in-Z. Thus, we could potentially get a different set of principles for each choice of *f* and Z.

These sets of principles-on-*f*-in-Z are not necessarily 'real' principles, or principles that are relevant to us, since principles that are only feasible on certain FCs may not meet whatever feasibility requirement there is on the truth of 'ought'-claims. We have not had an argument that any specific sharpening(s) of 'feasibility' are constraints on the truth of 'ought'-claims, or on 'ought'-claims of some specific kind, so we do not know which of the

sets of principles-on-*f*-in-Z are principles that really hold for us. The question that I will be engaged with in the third chapter is the question which of these sets of principles-on-*f*-in-Z we should be interested in. As I suggested above, the answer need not be just those that are principles that actually hold for us. It could be the case that, even though some FC *f* is not the relevant constraint for the truth of any moral principles (or 'ought'-claims), knowing what we *would* be required to do given *f* as the only constraint is useful for deciding what we *actually* ought to do (or conversely it could be that there are principles that *do* hold but that are not of interest for action-guidance). In a sense, theories that identify the correct principles-on-*f*-in-Z for each *f* and Z are all different possible types of theory. However, it seems likely that at least some of these types of theory (at least some of the sets of principles-on-*f*-in-Z) will not be of any real practical interest, and so political and moral philosophers should not devote time to identifying *these* sets of principles-on-*f*-in-Z. For instance, it might be thought that principles established only taking the laws of physics as constraints, or conversely principles established taking all of the facts of the world as constraints, are not of any practical interest. My argument will thus be about which of these possible types of theory are actually important or worthwhile.

I set out in this thesis to investigate feasibility critiques. One type of such critique rejects a moral theory as correct by saying that the principles it puts forward do not hold since their observance is not feasible. (Such a critique will obviously only be successful for those sorts of theories or principles that *are* constrained by feasibility of some kind). The sort that I will be interested in for the rest of this thesis rejects a moral theory instead as *interesting*. It says that we should not be interested in some moral theory because its observance is not feasible. Given the account I have given of the meaning of 'feasibility', it will not be enough simply to say that the observance of some theory is not feasible, since most likely on some sharpenings it will be, while on others it will not. A feasibility critique of this type, then, relies on the claim that the criticised principle is not a correct principle-

on-*f*-in-Z for some particular *f* and Z (or some particular *f*s and Zs), since its observance is not feasible on *f* in Z. This takes the simple form of a modus tollens: $\frac{O \rightarrow F}{\frac{\sim F}{\sim O}}$. Since the principles-on-*f*-in-Z are just those principles that we get when we *assume* that *f* is a hard constraint on moral theory, this just follows straight from the unfeasibility of the observance of some principle on the given FC. Such a critique combines this claim with an argument that the FC in question *is* relevant to us and that all interesting moral theories of the type in question *must* be feasible *on that FC*. (Alternatively, it could claim that a principle must be feasible to a certain degree (in the scalar sense of 'feasibility') in order to be interesting, i.e., that the lowest FC on which it is feasible must be below a certain point on the FC-scale and that it is not a correct principle-on-*f*-in-Z for any such FC). Such a critique obviously only works once it is accepted that some fact about the feasibility of a proposal on the given FC(s) is correct, that is, once it is accepted that the appropriate response to the above modus tollens argument is not to reject the $\sim F$ premise. This critique, then, relies on the claim that an interesting theory must be feasible on or below some particular FC, that is, that we should only be interested in certain sets of principles-on-*f*-in-Z. The argument that this is the case for some FC is what I will discuss in my third chapter. First, I will set up a framework for understanding how the different sets of principles-on-*f*-in-Z relate to the most fundamental principles, which will be useful in making that argument.

Relations between feasibility constraints and the most fundamental moral principles

The principles or moral 'ought'-claims that are of direct action-guiding use (i.e., those that tell us exactly what to do now) are likely to be specific to a certain, relatively narrow, factual context (that is, they will tell us what to do in the actual world and perhaps will hold for a reasonably narrow set of possible worlds beyond it). The most fundamental moral principles relevant to us are those principles that underlie our other principles at

the deepest level that there is. They may be entirely independent of the factual context (this is what Cohen argued[54]) or they may not, but either way, they will presumably hold across a wider range of factual contexts than those principles that tell us directly what to do now. One may get to a principle underlying some moral judgement by asking *why* the thing in question is good or bad, right or wrong, or whatever. I take it that this principle is not really more *fundamental* than the original judgement or principle if it is not more general in this way, that is, if it does not cover a wider range of factual contexts or possible worlds. Thus, the most fundamentally correct moral principles hold across a range of possible worlds. It is a matter of potential debate how wide the range of possible worlds over which they hold is. I leave it open whether they are true for all possible worlds or whether they only hold contingently on certain facts of the world, that is, they only hold for worlds identical to the actual world in certain respects. I will now set out a framework for understanding the relationship between these most fundamentally correct moral principles and the different sets of non-fundamental principles-on-*f*-in-Z (that is, the different sets of 'ought'-claims we would get when different FCs are taken as constraints on non-fundamental principles). This will thus be a framework for understanding how we get from the fundamentally correct moral principles to principles that have appropriately taken the relevant feasibility facts (those fixed by a chosen FC) into account.[55]

[54] Cohen (2003) and (2008, ch. 6). This is denied by, e.g., Pogge (2008).

[55] The framework I will set out is a framework for describing how moral principles for governing action relate to facts about feasibility. It will thus not say anything of interest to someone who denies that there are (general) principles of right action, such as a virtue ethicist might do. That is not to say, of course, that considerations of feasibility would not be of interest to such a theorist, but their role would require a very different treatment.

Consequentialist theories

A consequentialist theory says that one should produce the best outcomes out of those feasible and gives some sort of criteria for ranking outcomes.[56] A fully developed consequentialist theory should, in principle (though obviously not in practice), be able to give a complete ranking. This is not to say that consequentialist theories must be maximizing. Maximizing consequentialist theories take one value, or some weighted bundle of values, and ask us to maximize it. Such theories thus give very straightforward rankings. Outcomes are just ranked according to how far they realize the value, or weighted bundle of values, in question. However, outcomes may be ranked by a consequentialist theory in a more complex manner. A theory could give more complex principles for ranking outcomes (perhaps giving balancing principles to be observed differently in different circumstances) or even just a piecemeal ordering of states of affairs. In any case, to be complete, a consequentialist theory needs some way of saying which outcomes are best.[57]

Thus, the operation of feasibility constraints on consequentialist theories is very straightforward. They are like budget constraints in an economic optimization model. They put an upper limit on the best outcome in the ranking that is possible within the given constraint. Thus, when we select an FC on which to theorize (i.e., to act as a constraint on 'ought'-claims), it just rules out a number of the outcomes in the ranking;

[56] Such a formulation of consequentialism obviously builds feasibility considerations into the definition. I think, though, that any plausible formulation of consequentialism will do so. It is hard to see what the claim that we should produce the best consequences could mean without some restriction to the best consequences possible or feasible. This is often formulated with reference to *possibility* rather than *feasibility*. I do not equate the two, but I think that when they are distinguished it is *feasibility* and not possibility that is relevant to the formulation of consequentialism. Feasibility is about what outcomes are possible results of human action, whereas an outcome is possible if there is some possible world (however 'possible' is glossed) in which that outcome comes about. A plausible formulation of consequentialism must surely require only that we produce the best consequences *feasible*, since it is irrelevant to our action choices what outcomes are possible if they are not feasible.

[57] Of course, a consequentialist theory could just leave it indeterminate for certain pairs of outcomes which is better, but then it would be incomplete.

the theory says choose the highest remaining one. Thus, there is not really a question of a consequentialist theory being unfeasible on some FC, or too unfeasible (in the scalar sense). No FC could rule out a consequentialist principle as a possible principle-on-*f*-in-Z, since there is no FC at which it is not feasible to produce the best outcome *feasible*.[58] In other words, the sorts of feasibility critique detailed above cannot apply to a consequentialist theory. Feasibility cannot be used to defeat consequentialist theories, only to establish the limits on them. Thus, I do not think that the standard sort of over-demandingness objection to utilitarianism can be construed as a feasibility critique (at least in my sense). It is, rather, that act-utilitarianism imposes too great a loss on agents or that it ignores a personal prerogative that is morally important.[59] This is a moral critique (saying the theory ignores something morally important), not a feasibility critique.

What might seem like a feasibility critique of consequentialism is the critique that says that it is not feasible for people to internalize the consequentialist rule, that is, for people to be internally motivated by that rule. However, this is not a criticism of consequentialism as I have characterized it. A moral theory gives recommendations for action. It classifies actions (and perhaps also motivations) as right, wrong, permissible or impermissible. If a moral theory says 'Do the action, of those available, with the best consequences', it does not tell us to evaluate (or choose) actions according to the motivations behind them. Parfit notes that consequentialism *does* require us to cultivate the motivations (of those available) whose being had by us will have the best outcome.[60]

[58] Of course, this is only the case so long as the consequentialist principle 'produce the best outcomes feasible' leaves vague the term 'feasible', to be sharpened according to the FC that is being assumed to constrain moral theory. A consequentialist theory could instead specify a specific sharpening of 'feasible', in which case it *would* be possible to reject it for violating an assumed feasibility constraint, if the FC assumed to constrain moral theory was different to the one taken to sharpen 'feasible' in the formulation of the consequentialist theory.
[59] This is, I think, the sort of objection offered by Williams (Smart and Williams, 1973).
[60] Parfit (1984) 26

However, this does *not* require us to be motivated by producing the best consequences. People's inability to internalize the consequentialist rule is not as such a constraint on the feasibility of this moral theory. It is still feasible for people to do those actions required by the consequentialist principle, and it should not fail to be. If motivations are taken as a constraint on feasibility (that is, if something is unfeasible if people will not be motivated to do it), it is still feasible to observe the consequentialist principle. Since changing motivations is assumed to be unfeasible, then the best action *feasible* is the best action consistent with those same motivations that are held fixed. If, on the other hand, motivations are not taken as a constraint on feasibility, then it *is* feasible to do what people are not motivated to do. Of course, if it is unfeasible for people to be motivated by the consequentialist principle, then a theory requiring people to be *motivated* by that principle could be ruled out (assuming the appropriate FC). This would not, however, be a straightforward consequentialist theory. Consequentialism, as a theory of right, does not require the consequentialist principle to be adopted as a decision procedure.[61]

A general framework

Principles that do not reference feasibility and demand or rule out specific types of action, unlike consequentialist principles, *can* be ruled out by a particular FC once it is assumed as a constraint on moral theory. If a principle demands we perform some type of action that it is not feasible for us to perform on a given FC or demands that we not perform some type of action that it is not feasible for us not to perform on a given FC, then that principle cannot be a correct principle-on-*f*-in-Z for that FC. However, the case is in fact not that different for deontological and consequentialist theories. This is because a plausible *complete* deontological theory does not just consist in a set of constraints that

[61] This was recognized by the early utilitarians, Bentham (1961, Chap. IV, Sec. VI), Mill (1863, Chap. II, Par. 19) and Sidgwick (1907, 413) as well as argued for by philosophers more recently, e.g. Bales (1971) and Parfit (1984, Ch. 1).

may or may not be feasibly observable, but, like consequentialist theories, takes account of feasibility. The way the most fundamental moral principles relate to the correct principles-on-*f*-in-Z for some *f* and Z is the same for theories on which the most fundamental moral principles are deontological and on which they are consequentialist.[62]

Hamlin and Stemplowska identify a form of theory that they call 'theory of ideals'.[63] The purpose of this, according to them, is to 'identify, elucidate and clarify the nature of an ideal or ideals'. This includes both an element 'devoted to the identification and explication of individual ideals or principles' and another 'devoted to the issues arising from the multiplicity of ideals or principles (issues of commensurability, priority, trade-off, etc.)'.[64] They describe this form of theory asking us to imagine a graph plotting the realization of two (or more) values or principles against each other. The task of the theory of ideals then involves both specifying the axes (that is, identifying what the values and principles *are*) and then identifying the shape and position of the indifference curves (that is, identifying between which bundles of realization of different values and principles we are indifferent). Thus, on a simple model with only two values, say equality and security, this would involve analysing what these values are or what they involve and deciding how they should be balanced when there is a limit to how much we can achieve of each.

Now described in this way, the theory of ideals fits easily into a consequentialist optimizing framework. It specifies the values in terms of which a ranking (of outcomes) is to be made and the balancing principles that determine how these values interact to give a full ranking (given by the indifference curves). Once we have drawn these indifference curves and understood what the axes are, all that would remain would be to draw some feasibility frontiers and then to produce the best outcome given a chosen frontier.

[62] Wiens (forthcoming, 15) suggests something similar.
[63] Hamlin and Stemplowska (2012)
[64] Ibid. 53

However, this consequentialist form of theory is a subspecies of a wider form of theory that I will call ideal theory of principles. This is that theory which identifies the most fundamentally correct moral principles. Extending Hamlin and Stemplowska's description of the theory of ideals to this form of theory will help to see how both deontological and consequentialist fundamental principles relate to different sets of principles-on-*f*-in-Z.

The correct principles may be consequentialist or deontological in form. If they are consequentialist then the most basic correct principle would say 'produce the best consequences feasible' (or something like that). Beyond this, the task of ideal theory of principles would be to specify the values and the balancing principles that determine which consequences are better than others. This may involve specifying further principles since specifying a value may be done by specifying principles that: the more (or better) they are observed, the more the value is realized. If the correct principles are deontological, on the other hand, ideal theory of principles involves simply specifying the rules that should constrain action according to the most fundamental principles of morality. These rules may conflict with each other. That is, there may be possible states of affairs in which two of the correct moral principles require us to do conflicting things (i.e., in which it is not possible to observe both of the principles together). Thus the set of correct moral principles (if deontological) will need something like the second element in Hamlin and Stemplowska's theory of ideals, some sort of balancing principles. That is, it will need principles that specify: a) how to deal with conflicts between the 'first-order' correct principles; and b) that indicate how other principles should be modified if some principle is not observed. In other words, principles are needed to deal with the interactions between different principles. The former sort will presumably give some rationale for selecting certain possible states of affairs where two principles conflict in which one of these principles should override the other and perhaps certain other conflict states of affairs in which a new principle supersedes both of the conflicting principles. The

latter sort will, for any principles p_1 and p_2 that are not wholly independent of each other, indicate how p_1 should be modified or superseded in cases where p_2 is not, or only partially, observed. Thus, if the set of correct deontological principles is complete (including balancing principles that cover all possible states of affairs) it should give us a recipe for determining the correct principles-on-*f*-in-Z, for each *f* and Z. We simply take those individual principles and combinations of principles that cannot be observed consistently with *f* and use the balancing principles to produce a modified set of principles.

Now, one might, of course, doubt whether there are complete balancing principles in this sense. It might be thought that there are no general principles that tell us how to balance principles for *every* possible situation. There are just some intractable conflicts that can only be resolved using some sort of 'moral judgement'. If this is the case, then the implications for action of the correct moral principles will not be fully determinate. The full set of correct principles-on-*f*-in-Z for every *f* and Z may also not be determined by the generally correct moral principles. Whether or not this is the case is unclear. It seems likely, though, that intractable conflicts will at least not be universal and thus that the correct moral principles will at least give some guidance for action given particular FCs.

One might perhaps also object to this understanding of moral principles that the sort of balancing principles it involves require a consequentialist framework. To assume that deontological principles can be balanced in this way is to assume that they can be subsumed under a consequentialist framework. I think my description of the relation between fundamental principles and principles on different FCs does in a sense require a consequentialist framework, but only in a trivial sense. It does not turn deontic principles into consequentialist ones; they remain deontic constraints on action, not just standards by which to assess the goodness of outcomes (albeit not absolute constraints, but I think

deontologists would generally agree that most, if not all, deontic constraints are not absolute). It requires a consequentialist framework only in the trivial sense in which all deontological theories can be restated as consequentialist ones. As Hooker says, 'on some of the ways of conceptualizing "good consequences", every moral theory can be formulated as some form of act-consequentialism'.[65] A deontological theory with no consequentialist elements (that is, one that merely consists in a set of deontic rules, constraints on action) can be trivially made into a consequentialist theory by describing it as the theory that says 'produce the best consequences, where consequences are evaluated according to deontological constraints *x*, *y* and *z*'. This need not be aggregative or maximizing; it need not say 'maximize observance of constraint *x*'. If it did the latter it would make the theory act-consequentialist in a less trivial way; it would require agents to produce the greatest overall amount of observance of principle *x*, which would aggregate across all agents' observances. We can, though, have a trivially consequentialist theory that is agent-relative that says to each agent 'what *you* should do is produce the best consequences, where the best consequences are those consisting in *you* observing constraints *x*, *y* and *z*'. If the deontological theory in question allows us to give balancing principles for handling cases where the principles cannot all be observed, then that theory trivially restated in consequentialist form allows us to give a ranking of outcomes and thus to say that given some FC *f*, one should produce the best outcome consistent with *f*. If it is denied that fully general balancing principles are possible, then the consequentialist restatement of the theory will not give a complete ranking and so will be indeterminate as to the best action on certain FCs. The deontological theory in its ordinary form is indeterminate in this way too, however, so this does not imply that the consequentialist restatement is different from the original theory.

[65] Hooker (2009) 153

Chapter 3

I argued in the previous chapter that for every FC on the FC-scale there is a form of moral theory that we could conceivably carry out. My aim in this chapter is to ask which sets of principles-on-*f*-in-Z moral and political philosophers should be interested in, or which FCs they should be interested in taking as constraints on their theory, insofar as their aim is to provide moral guidance for action. I will not give a positive answer of any precision, but will rather argue that they are not as restricted as has been claimed. It is sometimes thought in political philosophy that theory should be constrained by reasonably low, 'realistic' FCs, since if it is not it will succeed only in giving recommendations for action for non-actual possible worlds in which the facts are different in the way that higher, less restrictive, FCs allow them to be, and these cannot be relevant for guiding action. Given the account of feasibility that I have given, this point cannot be argued simply by claiming that 'unrealistic' theory is not of interest because what it calls for is not feasible. It *is* feasible on some FCs. Thus, it must instead be argued that certain sets of principles-on-*f*-in-Z are not of interest. The claim would be that when we allow too large a range of facts not to constrain our theory, that theory ceases to be of practical use. I want to argue in this chapter that this is not the case, that relatively unrealistic theory plays an important role for action guidance in the actual world.[66] I will use 'unrealistic theory' to mean theory constrained only by relatively high, expansive FCs (which are those that allow a wider range of facts to vary and consequently a wider range of proposals to be feasible) and 'realistic theory' to mean theory constrained by lower, more restrictive FCs. Thus, the argument of this chapter takes place against the background of this fairly common criticism of unrealistic theory that it is not action-guiding for the actual world. This criticism does not involve the claim that some proposed principle cannot be a correct

[66] I will remain vague about how high I mean by 'relatively high'. It may be that there are *some* very high FCs that are too expansive to be of interest.

moral principle unless it is action-guiding.[67] Rather, as Adam Swift notes, the claim of the critics 'is more helpfully conceived as a normative claim about what kind of theoretical work is important or valuable than as an attempt to identify the proper purpose of political philosophy or to specify what should and should not qualify as a theory of justice'.[68] I will accept, at least for the sake of argument, that *direct* action-guidance (that tells us exactly what we *actually* ought to do *now*) is necessary and constrained by relatively low (realistic) FCs. I will argue that, nevertheless, we ought also to be interested in what we should do given more expansive feasibility constraints: unrealistic theory is interesting and important for establishing these *directly* action-guiding 'ought'-claims, that is, it can be *indirectly* action-guiding.

In the literature this action-guidance criticism is a standard objection to *ideal theory*.[69] The distinction between ideal and non-ideal theory was introduced by Rawls, and he made the distinction in terms of theory that assumes full compliance and theory that does not.[70] Full compliance theory is not, though, the only sort of unrealistic theory that could be subject to the criticisms levelled against ideal theory, it is just one sort. Unrealistic theory is at least in general ideal in a sense since the principles identified given any particular FC will be those that demand what is morally best given the FC in question. Thus, since more expansive FCs generally allow for a wider range of possibilities, presumably the principles established on higher FCs will be more ideal in this sense. In any case, regardless of how we understand 'ideal theory' and what kind of theory this objection has been aimed at in the literature, one might think that the objection applies

[67] In a sense this is true, since moral principles are about what actions or sorts of action we should (or should not) perform. However, the sort of action-guidance demanded by the criticism is stronger, it demands *direct* action-guidance, which tells us what, concretely, we should do *now*.
[68] Swift (2008) 368
[69] This is the objection discussed by Valentini (2009). The criticisms of ideal theory presented by, e.g. Sen (2006) and (2009), Goodin (1995) and (2012), Farrelly (2007), Phillips (1995) and Miller (2008) are all variants of this objection (or include variants of this objection).
[70] Rawls (1999a) 8

to unrealistic theory because it provides recommendations only for imaginary worlds in which many more proposals are feasible than in the real world, and thus fails to provide guidance for action. For the purposes of this chapter then I will use 'ideal theory' to refer to unrealistic theory. Thus, whether or not a theory is ideal will be a matter of degree. I do not mean to suggest, however, that this is the only appropriate use of the term.

The problem of second-best

One basic, and important, thought in response to this action-guidance objection to ideal, or unrealistic, theory is that we need ideal theory to get an idea of what we are aiming for or to act as a benchmark by which to measure the moral acceptability or goodness of different options. (In any case, even if we do not *need* ideal theory, it may be *useful* for action guidance in these ways). At least certain sorts of unrealistic theory can be relevant to guiding action in the actual world, since when we know what we would be required to do when a wider range of the current facts can vary, we will have an idea of what sorts of outcomes we should direct our action towards achieving. Though our action will be constrained here and now (in the short term) by low FCs, we can choose actions within these low FCs with an awareness of what would be better and thus what we should strive towards. Theorising at relatively high FCs tells us what is 'ideal', or relatively ideal. This, it might be thought, gives us action guidance for what we should do, even though the theory is arrived at assuming that certain facts can change that it is unrealistic to think could change, because we should just attempt to get as close as possible to doing what we should do ideally. If, say, what we should do when we allow human motivations and so on to vary is achieve perfect equality of welfare, then when human motivations and so on are not variable and we cannot achieve such equality, what we ought to do, the thought goes, is get as close as possible to this ideal.

Furthermore, deeper, more fundamental principles are ones that hold more generally, that is, for a wider range of possible worlds and not just for the current specific circumstances. The more fundamental the principles, presumably the higher the FC that constrains them will be. I have remained agnostic as to whether or not the *most* fundamental principles are constrained by feasibility at all. However, even if they are not entirely FC-free, they presumably are constrained at most only by a relatively high FC. If they are constrained by feasibility facts, that is, they will be constrained by fairly deep and permanent facts about the world such as the basic facts of human nature. More fundamental principles underlie the more context-specific principles that apply to us as well as the specific (and concrete) institutional choices we ought to take or actions we ought to perform. Thus (or, at least, so the thought goes), it will be useful in deciding which concrete actions to take or policies to adopt to determine some of the more fundamentally correct principles that underlie these decisions. That is, it will be helpful to identify the principles that are the reasons that we ought to choose some actions or policies and not others. Since the theory that identifies these principles is constrained only by at least a somewhat high FC, this means that unrealistic theory can play an important role in delivering action guidance. This is a variant of the idea that unrealistic theory can be useful in telling us what we ought to be aiming for since the deeper principles are what a morally motivated choice of action or policy is aiming to adhere to. It is because it realises such principles that a particular action or policy is the one that ought to be chosen. They are the standards by which actions and policies should be judged. Thus, the basic response on behalf of unrealistic theory is that it can provide action guidance by giving us a guide as to what we should aim for. Different types of unrealistic theory can do this in different ways.

Rawls's defence of ideal theory is an argument of this sort. 'Until the ideal is identified', he says, 'nonideal theory lacks an objective, an aim, by reference to which its queries can

be answered'.[71] Other variations on this argument are present in various other contributions to this debate.[72]

This, then, seems to be a fairly straightforward and strong defence of ideal theory as action-guiding. However, it has been noticed that there is a result in economic theory that applies generally to the relation between optimal outcomes and 'second best' outcomes, which seems to pose a problem for this argument. In 1956 Richard Lipsey and Kelvin Lancaster proved a theorem in economics that they called 'the General Theory of Second Best'.[73] The idea is that if a Pareto optimal outcome consists in the fulfilment of a number of 'Paretian conditions', then

> given that one of the Paretian optimum conditions cannot be fulfilled, then an optimum situation can be achieved only by departing from all the other Paretian conditions ... Specifically, it is *not* true that a situation in which more, but not all, of the optimum conditions are fulfilled is necessarily, or is even likely to be, superior to a situation in which fewer are fulfilled.[74]

Robert Goodin, notably, has discussed the application of the theory of second best (TSB) to ideal theory in political philosophy.[75] Goodin notes that the very strong conclusion Lipsey and Lancaster arrive at (that an optimum situation can be achieved *only* by departing from *all* the other Paretian conditions) stems from certain assumptions they made. However, the latter conclusion they make in the quote above holds more generally, he suggests. The second-best state of affairs is not *necessarily* identical to the first in any respect.[76] The idea, then, is that if our ideal theory calls for something (a state of affairs, a set of institutions, a set of principles) that has several features, but the

[71] Rawls (1999b) 90
[72] For instance, Stemplowska (2008) and Gilabert (2012).
[73] Lipsey and Lancaster (1956)
[74] Ibid. 11-12
[75] Goodin (2012) and (1995). Michael Phillips (1985) also raised a similar problem for ideal theory.
[76] Goodin (2012) 157

constraints of the world prevent all of these features from being achieved *together*, the second-best alternative is not necessarily going to be just the one that is closest to the first-best in the greatest number of these features. The alternative that changes the least features of the ideal may actually be worse by the lights that led us to choose the first-best than one that changes more. Thus, if we take the best principles or outcomes given some reasonably high FC as an ideal, it will not necessarily be the case that the best principles or outcomes given a lower FC will resemble them. Thus, if a high FC allows facts to change that it is very unrealistic to think could change, then, the thought goes, a theory constrained only by this high FC will not be much use as a target or standard for guiding action, since there is no guarantee that what we should do given a more realistic FC will resemble the target. Thus, it is not obvious how knowing what the target is could help us in judging principles or policies in the real world.

Goodin gives the analogy of a choice of car. Suppose my ideal car would have three features: it would be silver, new and a Rolls Royce. Suppose now that such a car is unavailable but two others are. One is a week-old black Jaguar and the other is a new, silver Toyota. The latter has two of the three features of my ideal car, while the former has none. However, it is likely that I would in fact prefer the Jaguar and not the Toyota.[77] Since unrealistic theory involves assuming away certain constraints that will constrain our actions in the real world, these latter are bound to be constrained by things that did not constrain the unrealistic theory. Thus, it will often not be feasible, given more realistic FCs, to do what the principles-on-*f*-in-Z for high FCs require us to do. Thus we are obliged to settle for a second-best alternative. The TSB appears to show that ideal theory will not necessarily be a useful guide to what that alternative is. This, then, suggests that identifying a target may in fact not be action guiding at all.

[77] Goodin (1995) 53 and Goodin (2012) 157

Response to the problem of second best

I think that the TSB does pose a problem for unrealistic theory that must be taken seriously. However, while there are important conclusions that *do* need to be drawn from this problem, I do not think it succeeds in showing that unrealistic theory is not an important enterprise for political philosophers to engage in. I will argue for this claim below. One response to the TSB is present in Simmons's (Rawlsian) defence of ideal theory. Simmons argues that 'a good policy in nonideal theory is good only as transitionally just – that is, only as a morally permissible part of a feasible overall program to achieve perfect justice, as a policy that puts us in an improved position to reach that ultimate goal'.[78] Thus, he argues that we should not simply be aiming to choose the most just policy or action of those feasible in our nonideal circumstances (i.e., given a realistic FC). If this latter were what we should be doing, then the TSB might seem to show that ideal (or unrealistic) theory is not much use, since knowing what the ideal principle or outcome is does not necessarily tell us anything about what the best principle or outcome is given some lower (more realistic) FC. However, Simmons thinks, the aim of nonideal theory should be rather to identify those principles or outcomes, consistent with a realistic FC, whose observance or obtaining will get us closest to achieving the ideal (perfect justice). This may *not* be what is most just, since we could have a case where one step backwards allows us to make two steps forward (an option that is worse judged by itself in terms of the fundamental principles could get us closer to the ideal). In other words, actions and policies are not to be judged by their comparative adherence to fundamental moral principles, but rather by the likelihood with which they will lead us to moral perfection. If our choices of actions and policies (and thus of principles for the nonideal world) should be guided, as Simmons says, by the aim of transition to the ideal, then it seems clear that it will be important to determine what that ideal is. Even though

[78] Simmons (2010) 22

in a sense something like the TSB may hold even for this aim, in that the actions or policies that achieve the most transitionally may not be those that most resemble the ideal, it seems like we can have no way of determining what is transitionally best without determining what the ideal is that we are aiming to get to.

However, though there is something in this (transition is an important element of evaluating proposals for action), one might object to this picture that it gives too much weight to the ideal. If there is a constraint preventing us from getting to the ideal which is so strong (and permanent) a constraint that it is reasonable to think we will *never* reach the ideal, it does not seem like we should opt for an action that would put us on the path that would lead to the ideal if that constraint were removed, where this involves foregoing an action that would be morally better *given* the constraint. Sometimes, it seems, we should opt for what is the best possible *given* certain feasibility constraints, rather than what would form part of the best strategy for achieving something better if certain constraints were removed. We should not always focus all our efforts on making the best outcome more likely, when that outcome is very unlikely and we could achieve something much better in the short term while making the best outcome less likely.

However, I think there is another response available to the TSB. I think that once we distinguish two different sorts of unrealistic theory, it becomes clear that the TSB is not a general problem for the ability of unrealistic theory to provide action guidance for the nonideal world. On one way of understanding the task of ideal theory, the TSB is very straightforwardly a problem. This understanding fits well with Goodin's car analogy above. It involves seeing ideal theory as doing something similar to choosing one's ideal car. My ideal car is a (possibly nonexistent) car that is the best car I could imagine. The task of identifying one's ideal car is essentially a task of *design*. One designs a car, exactly the way one would like it to be. If we think of ideal theory as just like this, except for

society instead of for a car, then the TSB poses a problem exactly analogously to how it does in the case of the car.[79] Unrealistic theory is then the task of high-FC *institutional design*. The term 'institutional design' is potentially misleading since, as well as institutions, this type of theory could involve designing policies or specific actions.[80] I will keep the term, though, since it is used in the literature, but I do not mean to exclude these other objects of design. In unrealistic institutional design we specify, as exactly as possible, how society should be, given a high feasibility constraint (i.e., given that many/most/all options are feasible), designing the institutions that would create the best society possible. This society then, like the ideal car, will have a number of attributes. Common sense might say that in order to achieve the best society we can within certain non-ideal constraints, we should create a society that instantiates as many as possible of the attributes of the ideal society to as great an extent as possible. However, the TSB appears to show that this is not the case. We cannot assume at all that the best thing to do, given the unachievability of the ideal society, is to create a society that instantiates more rather than less of the attributes of that ideal society. If this is the case, then this sort of ideal theory seems to give us very little guidance as to what to do in a world in which the fully ideal society is not achievable. I will, though, argue below that there is nevertheless an important role for at least reasonably unrealistic institutional design.

Institutional design is not, though, the only form of unrealistic theory. In chapter 2 I discussed what I called 'ideal theory of principles'. This is that theory that attempts to identify the most fundamentally correct moral principles, and theory of principles generally is theory that attempts to identify correct moral principles. Earlier in this

[79] Phillips (1985) criticism of ideal theory also seems to involve understanding ideal theory as being something like institutional design.
[80] Robeyns (2008) prefers the term 'action design' for this reason.

chapter I suggested that this theory will be constrained, if at all, by a reasonably high FC.[81] Hamlin and Stemplowska describe the theory of ideals as involving specifying the axes on a graph plotting the realisation of two or more values against each other and specifying the shape and position of the indifference curves.[82] Theory of principles is analogous to theory of ideals, only it covers principles generally, thus allowing for non-consequentialist moral theories. They say that the role of institutional design comes into play once we add feasibility frontiers to such a graph. We should aim to be at the point at which the highest indifference curve touches the relevant feasibility frontier. The task of institutional design is to design institutions that will achieve this. It will not be possible to draw straightforward indifference curves for non-consequentialist theories but the basic relationship is the same. Institutional design seeks to design the actions and institutions that will best achieve the most fundamentally correct moral principles (as identified by the theory of principles) given some choice of feasibility constraint.

The way Goodin presents the problem could suggest that it is meant to apply to unrealistic theory of principles.[83] The TSB shows that if, say, theory of principles calls for the maximisation of a number of values, in a situation in which they cannot all be realised, the best way to balance the values against each other is *not* necessarily to maximise as many as possible. If all that ideal theory of principles told us was what values we should promote, then the TSB would leave that ideal theory more or less impotent to guide non-ideal theory for the actual world.

However, in chapter 2 I argued that theory of principles in general (whether for consequentialist or deontological principles) will need balancing principles, which tell us

[81] Estlund (2014) argues persuasively that such principles, though they can be constrained by certain facts about what people *can* do, are not constrained by facts about what people are *likely* to do. This could be interpreted as the view that they are not constrained by lower FCs that hold fixed facts about human motivations and the like.
[82] Hamlin and Stemplowska (2012) 53-4
[83] Goodin (1995)

how to deal with cases of conflict between first-order principles and cases where first-order principles cannot be, or are not, observed (or cases where the first-order principles cannot be observed together).[84] I also argued that a complete theory of principles including balancing principles would provide a recipe for determining the correct principles-on-*f*-in-Z, for each *f* and Z. Complete balancing principles should show, for any case where the fundamental principles cannot all be straightforwardly observed, how they should be modified or superseded. Thus, for any given feasibility constraint that makes certain outcomes impossible, it should give us a new set of principles that it is feasible to observe. Thus, the TSB does not seem to be a problem for this form of theory. It does not just provide us with a particular state of affairs that we ought to approximate to, or a set of values to realise as much as possible. When designing balancing principles is included as part of the task of the theory of principles, it does provide us with action guidance for nonideal circumstances (i.e., where lower FCs are operative). Theory of principles for car choice would involve determining what the principles are that make cars good and some cars better than others, and this would involve determining second-order principles for balancing these principles or values. Thus, when we are faced with a car choice with a feasibility constraint, the theory of car principles will be useful for selecting which car is best; we choose the car that achieves the best balance of principles as determined by our theory of car principles. Thus, unrealistic theory of principles must take account of the TSB, in that in specifying balancing principles, one needs to be aware of the fact that the next best alternative is not always the one alike in the most features to the best. However, the TSB does not make this sort of ideal theory irrelevant to nonideal decisions.[85]

[84] Following on from how Hamlin and Stemplowska (2012) characterise the theory of ideals.
[85] Adam Swift (2008) argues against Goodin that we need to know not only what the ideal car looks like but also why, so we know what to value in a car in general (377). This is, Swift thinks, the value of ideal theory; it lies not simply in designing an ideal car (or society), but in identifying *why*

I should note that this sort of defence of ideal theory against the TSB does not involve saying that political philosophy should, in practice, proceed by first establishing principles and balancing principles and only then applying them. The point is just to see that the TSB does not, as such, make ideal theory irrelevant to action constrained by the feasibility constraints of the actual world. A part of the task of ideal theory makes the TSB no longer problematic. Even though it is practically impossible to carry out the task of ideal theory fully, that does not make that ideal theory that we *can* do irrelevant to real world action. If ideal theory cannot be applied without balancing principles, all this shows is that we need to do more ideal theory, not that ideal theory is useless.

The problem of second best for institutional design

Now, one might accept all of my arguments so far but argue that the TSB nevertheless still does pose a problem for political philosophy. Even if it does not make ideal theory of principles redundant, unable to guide us in implementing its own recommendations in a nonideal world, it does create a puzzle about this implementation. The puzzle is not about the move from theory of principles to institutional design, but rather about how to do institutional design itself. Institutional design is an important task: to implement the fundamental principles, it will be necessary to design the social institutions or actions that will best realise the requirements of those principles. This design can be carried out within the constraints of a whole range of different FCs. That is, there is a whole range of 'realisticness' with which it can be done. There is not one single FC that is obviously the only appropriate constraint for institutional design. Choosing more expansive FCs is similar to designing actions with a longer-term in mind. There is no obvious right answer to the question how constrained by short-term circumstances one ought to be. For

---

the ideal is ideal. The idea behind this seems to be similar to what I have argued, since it seems like identifying why the ideal car is ideal will involve identifying the principles that underlie our valuing the car.

example, ought we to take the current motivations and preferences of people as constraints on our institutional design or not? It is possible to design the actions that would be best *if* a given FC is assumed as a constraint. However, it is not obvious which of these designs are useful in guiding our decision of which actions to actually opt for.

The TSB, then, might seem to pose a problem for our choice of which FC to use for institutional design (insofar as we are aiming to carry out institutional design on those FCs that allow our institutional design to guide our real-world action). If, say, the best possible institutions given constraint f_1 are X, Y and Z, the TSB shows us that we cannot assume that the best possible institutions given some lower (that is, less expansive) constraint f_2 will resemble X, Y and Z in any of their attributes. If we could assume that, then the choice of FC for institutional design should not matter that much, since we could just design institutions for some relatively high FC and assume that more short-term action (i.e., on a lower FC) ought to be geared towards institutions more or less similar to what we designed on the higher FC. We could, for example, design the best institutions not taking people's motivations as a constraint, and then assume that when they *are* a constraint, we should just do whatever most closely resembles the institutions we have designed. But given that we cannot do this we have something of a dilemma. On the one hand we could just focus on designing the best possible institutions for a relatively low, realistic FC. Since we want to guide action in the short-term this might seem to be what we should do. On the other hand, however, we want to be more demanding than this. We cannot be exclusively restricted to short-term FCs, since this is to ignore what Gilabert calls 'dynamic duties'.[86] We may have a dynamic duty if it is not now possible to do *x*, but it would be possible to do *x* if we did *y* and we *can* now do *y*. If *x* is something we should do and it is worth the cost of doing *y*, then we should do *y*. Because of this, Gilabert thinks, political philosophy should adopt a 'transitional standpoint', which 'focuses on the

[86] Gilabert (2011, 59ff) and (2012, 47ff)

identification of dynamic trajectories of political action, which set into motion a sequence of political reforms passing through successive thresholds of feasibility'.[87] Thus, we might prefer to design institutions on a higher FC in order to avoid excessive conservatism and to get the best institutions allowing for some greater changes to be made.[88] However, the TSB shows us that focusing on the long-term in this way may lead us to do things in the short-term that are not in fact the best things to do in the short-term. Thus, pursuing what is desirable given low FCs may prevent us from getting to better states of affairs that we would design on higher FCs. On the other hand, focusing on higher, more unrealistic FCs might lead us to end up with worse outcomes than we could have had.

This is certainly a dilemma and it is a difficulty for institutional design. I do not have an easy way out of this difficulty, but we should note that it is not a *theoretical* problem. If we have a complete ideal theory of principles and all the necessary social scientific knowledge, it should be possible to design the best possible actions given *any* FC. We should then be able to weigh up changes aimed at expanding the FC, that is, making new things possible, according to how much better the actions they would make possible are than the best that are possible on a lower FC and how much worse the short-term actions/institutions they would require are than the best actions/institutions possible on the lower FC as well as how likely those short-term actions are to succeed. There should (again, theoretically) be some function of the desirability of the best outcomes on different FCs as well as the expansiveness of the different FCs (that is, how difficult it would be to achieve the best actions on each FC) that could enable us to decide when opting for an action or institution that is not the best possible on some FC is the best choice to take given the better institutions/actions that it will make possible. Sometimes we should pursue the best institutions or actions on high FCs even if doing so involves

[87] Gilabert (2008) 411
[88] Valentini notes that the lower the FC, 'the more [the theory] will appear to offer an uncritical defence of the *status quo*' (2012) 659.

choosing actions/institutions that are not the best in the short-term. Other times, the badness of the short-term actions/institutions required to make better ones possible in the long-term will outweigh the latter's 'betterness', and instead we should opt for the actions/institutions that are the best on a lower FC. On the basis of such a function, it should be possible to determine which FC is the optimum one within which to design actions/institutions that achieve the best balance between long-term and short-term desirability. In practical terms, it may not be humanly possible to calculate this function very well. This, though, is just to realise that the task of political philosophy is an immense and difficult one. This is, in a sense, what political philosophy should aspire to. In practice, political philosophers should just do their best to balance the requirements of institutional design on different FCs and, using intuition, to come up with recommendations for action, for which the short term losses are more or less balanced by the long-term gains. What is important is that we should not just focus on institutional design given *one* FC, but rather within many different ones, since, given the TSB, only in the latter case can we achieve a proper balance between short-term feasibility and long-term demandingness.

Jonathan Wolff argues that, at least sometimes, political philosophers should not operate according to Rousseau's famous dictum, taking men 'as they are and laws as they might be', but rather also should take laws as they are.[89] He says that 'it is one thing to set up laws for an ideal society of the imagination, but the task in hand is to deal with the world we have. Inevitably, then, for policy reasons what needs to be discussed is not ideal law and regulation, but change to existing law'. Of course, this point is about institutional design, and as Wolff admits, without 'speculation about ideals ... there would be nothing

[89] Wolff (2011) 78; Rousseau (1996): 'en prenant les hommes tels qu'ils sont, et les lois telles qu'elles peuvent être' (45). Rousseau does not say that political philosophy *should* operate according to this dictum; he only announces that he *will* do so.

to inspire or direct change'.[90] We need ideal theory of principles in order to have an idea of what makes the laws or policies we design for the real world good or bad. However, when it comes to institutional design, I think it is right that the task of political philosophy is to 'deal with the world we have' (at least when the aim of political philosophy is to engage with and guide real world politics, rather than just to form a part of an intellectual exercise). It is not obvious, though, what exactly this requires of institutional design in terms of feasibility constraints. Since it is not obvious that there is any single FC that constrains what we can be required to do, or what we ought to do (in the domain of public policy or otherwise), as I have argued, theory carried out given a variety of different FCs will be useful for guiding action. To achieve a proper balance between short-term conservatism and long-term aspiration, it is important to consider what actions and institutions *would* be morally best given relatively high FCs *as well as* given lower FCs. The importance of dealing with the world as it is does not mean that institutional design ought to be entirely focused on design that takes the current state of the world to be mostly fixed. (Of course, there is a limit to this argument; though it is important to know what is best given relatively high FCs, i.e., allowing a wider range of facts of the world to change, there will presumably be some FCs that are *too* unrealistic to be of much use for this purpose).

There is an additional reason why unrealistic institutional design may be of use for action guidance. It is a less direct reason (and its importance is perhaps less: it gives a reason why high-FC institutional design may be worthwhile, but does not so clearly give a reason that it *ought* to be done) but it could vindicate institutional design carried out on even higher FCs than my main argument would be likely to support. This is a way in which it could be helpful as a *heuristic*. The most fundamental principles relevant to us, which I have argued it is important to theorise about, hold across a range of possible worlds. One

[90] Wolff (2011) 79

way of testing whether some principle is one we are fundamentally committed to, or whether there is rather some other more fundamental principle that explains its intuitiveness in the current case, is by asking whether the principle seems to hold when we change some facts. That is, we conceive of an imaginary situation that we think is within the range of possible worlds that our fundamental principles ought to hold for and we ask whether the principle in question seems to hold in this situation. If it does not, we discover that we are not really committed to that principle fundamentally, but only insofar as it realises some other principle. For example, Parfit's imaginary divided world where two halves of the world's population are unaware of each other's existence serves this purpose. If we prefer a situation where there is equality in each half but not between the two to one in which the two halves are equal but there is a lower aggregate utility to the former situation, then we are not fundamentally committed to a principle of equality. If on the other hand, we prefer the second world, we may be.[91]

To identify the most fundamentally correct moral principles, then, we need to consider non-actual possible worlds, in order to see what our intuitions about them would be. The best method for moral and political philosophy is often thought to involve reflective equilibrium, where we seek to bring our intuitions about particular cases into balance with our intuitions about general principles, rejecting intuitions on either side as seems appropriate when they come into conflict. Our fundamental principles are what underlie our judgements about institutional design; when we design institutions and actions we design those that best realise these principles. If what we think are our most fundamental principles dictate an institutional design for some factual situation that we find hard to accept, then, if we are more committed to this judgement about institutional design than we are to the judgements that led us to the fundamental principles we thought we were committed to, we may need to revise the fundamental principles. It is important, though,

[91] Thought experiment from Parfit (1997) 206.

that because the most fundamental principles should hold across a range of possible worlds, we need to test them not just by the institutional design that they dictate for the actual world, but also by the institutional designs that they would dictate for other counterfactual worlds. Thus, it can be useful to carry out institutional design for high FCs (as well as for worlds in which the factual base, i.e., the Z, is different) as a heuristic for our theory of principles.

Sen's objection

There is what may seem to be an important objection to my defence of unrealistic theory in Amartya Sen's objection to 'transcendental theory'.[92] Sen argues that what are needed for making real-world decisions are comparative judgements. That is, if some form of moral theory is to be action-guiding it needs to help us to make comparative judgements. In the real world what we need to be able to do is select between available alternatives and so we need to be able to compare these to decide which one we morally ought to go for.[93] Given what I have argued about feasibility it is not clear what are the 'available alternatives', but the idea is presumably that what is important is to compare the different proposals that are all feasible on some reasonably low (realistic) FC. The claim is that any other type of theory, though it may be intellectually interesting in its own right, is only of practical relevance to real-world action insofar as it contributes to this task of comparing available options. Let us accept that this is correct. It seems just to be a way of understanding the action guidance requirement.

Sen argues that what he calls 'transcendental theory' is neither necessary nor sufficient for making comparative judgements. In effect, transcendental theory, he thinks, is redundant for real-world decisions. He characterises the transcendental approach as

[92] Sen (2006) and (2009)
[93] Sen's discussion is about justice, but I assume that there is nothing about his objection that is specific to the ideal of justice and does not apply equally to other types of moral principles.

'focusing ... on identifying perfectly just societal arrangements'.[94] When extended beyond the ideal of justice to moral theory more generally, presumably 'transcendental theory' refers to those theories that seek to identify morally perfect societal arrangements. Since the only things that prevent us from achieving perfection are feasibility constraints, theory conducted on the very highest FC (i.e., where nothing is a constraint on feasibility) will be associated with perfection. Institutional design conducted on the highest FC would design the actions and institutions that would make society morally perfect and theory of principles conducted on this FC would identify the principles whose observance would make society morally perfect. Sen's objection, though, seems similarly strong against any theories conducted, though not on the very highest FC, assuming FCs that are deemed outlandishly high. Any moral theory seeks to identify the best institutional designs possible or the best principles observable on its chosen FC. If this FC makes things possible which it is outlandish to assume are possible, then the associated theory could plausibly be open to a criticism similar to Sen's. What it is outlandish to assume is possible is obviously an open question.

The notion of 'transcendental theory', however, is ambiguous between an institutional design interpretation and a theory of principles interpretation. Sen's own words, describing it as being about 'identifying perfectly just societal arrangements', seem to suggest the former interpretation. However, interpreted this way, the objection seems to become less interesting. Not many theorists have been concerned to design the best concrete institutions and actions for a *perfect* society where there are no feasibility constraints. If this sort of theory is what Sen is criticising, he may well be right that it is redundant in practical terms. However, a criticism of it also seems somewhat redundant. Unrealistic institutional design (that is not high*est*-FC institutional design) could also obviously be open to a redundancy criticism. I have argued above, though, that at least

[94] Sen (2006) 216

reasonably unrealistic institutional design is important and relevant for action-guidance. There still must be some point along the FC-scale (well below the top) above which institutional design ceases to be of much interest (except perhaps as a heuristic for identifying more general principles). A Sen-type criticism, though, does not help us to identify how high up the scale this

point comes. Sen's objection seems more interesting to me if we interpret 'transcendental theory' as referring to what I have called ideal theory of principles.[95] Theorists have often taken this sort of theory to be important and useful, perhaps a prerequisite, for comparing different possible institutional designs in moral terms. If it turns out that attempting to identify the more fundamental principles that underlie these comparative judgements is in fact redundant, then that is an important result. I have argued that these principles *do* underlie moral choices of institutional design and choices of principles-on-*f*-in-Z for any given *f* and Z. However, it might be claimed that even so, we do not need to know what these principles are in order to make comparative moral judgements between different possible institutional designs, and that pursuing the question of what the fundamentally correct moral principles are will not, in fact, help us make actual moral progress (choosing institutional designs that are morally better than those we have).

Sen argues that transcendental theory is neither necessary nor sufficient for making comparative judgements. I will focus on the claim that it is not necessary. Sen argues that disagreement about ideal theories of justice is deep-seated and intransigent. He suggests that debates about which ideal theories are correct gets nowhere and is ultimately futile. However, Sen argues, despite the intransigence of these disagreements in ideal theories,

[95] This interpretation does not seem to chime well with what Sen actually says, since he talks about societal arrangements and ranking alternatives. Thus, I do not want to suggest this as an exegetically accurate interpretation of Sen's argument. It is perhaps better, instead, to think of it as just another objection very closely based on Sen's.

it is perfectly possible to make progress in comparative theory (and thus in practice). There are certain judgements for which all of the plausible ideal theories of justice overlap in their implications. For instance, he claims, proponents of all of the principal theories of justice could agree that continuing famines in a world of prosperity are unjust. Thus, there is no need to seek the answers to the deeper questions in ideal theory, we can just proceed with comparative judgements by seeking overlap, that is, those comparisons that all can agree on.

However, I do not think that such an argument could go very far in showing that ideal theory of principles is redundant (in general) for the purposes of action-guidance. The role of political philosophy is not just to identify changes that everyone already agrees would be improvements. To some extent this ought not really to be a task at all. If everyone agrees that some proposed change would be a moral improvement over the status quo, then we ought to implement it. Obviously, there could be political or institutional obstacles to doing so, but overcoming these is not in itself a philosophical task. Of course, there may be changes that *if proposed* everyone would agree would be improvements, but that have not yet been suggested or thought of. Perhaps the task of coming up with, or designing, these is a philosophical task. It seems that it could equally well be the task of social scientists, but I am not concerned here with how the labour should be divided up. Nevertheless, even if this is something that political philosophers should be doing, it is surely not *the whole* task of political philosophy. There are a great number of comparisons of proposals where there would be significant *disagreement* over whether some alternative is better (morally) than another. Robert Jubb makes this point in relation to one of Sen's examples of a case that is supposed to be beyond disagreement:

> Sen claims "instituting a system of public health insurance in the United States that does not leave tens of millions of Americans without any guarantee of medical attention at all [would be] an advancement of justice" [Sen (2006) 217]. That is a controversial claim, with which – judging by their willingness to vote for candidates who actively campaign against it – apparently tens of millions of Americans disagree.[96]

There is thus a need for political philosophy to develop tools to identify which possible institutional designs are better in cases where it is not already obvious, or where there is disagreement. This, I think, is the primary task of political philosophy, attempting to resolve disagreements through rational argument. To do this we will need principles to say why one institutional design is better than another. Though there may be fairly deeply entrenched disagreement about these fundamental principles and making progress on them is no doubt difficult, to leave it aside entirely is just to put aside the difficult questions. If we do not ask the difficult questions we will be severely limited in the improvements we can identify.

Adam Swift argues against Sen that 'as long as philosophers can tell us *why* the ideal would be ideal, and not simply *that* it is, much of what they actually do when they do "ideal theory" is likely to help with the evaluation of options within the feasible set'.[97] With the aid of my distinction between institutional design and theory of principles we can see that what Swift is getting at here in fact relates closely to the two different points that I have made above. The first is my point about theory of principles. Identifying the most fundamentally correct principles will tell us *why* some alternative is better than

[96] Jubb (2012) 239. Robeyns (2012) also argues that 'many cases of injustice are complex and often subtle, and therefore more difficult to identify and analyse as a case of injustice than cases of basic injustice' (160).

[97] Swift (2008) 365. Similarly, Mason (2004) argues, 'theorists will need to understand *how* the considerations they cite as reasons for thinking that some set of arrangements are ideal under present historical circumstances justify that conclusion' (254).

another. They are the principles that underlie any comparative judgements, and so identifying them ought to be useful (and sometimes is necessary) for making these.

The second point relates closely to my defence of unrealistic institutional design as a heuristic. So long as we do not *just* identify the ideal actions and institutions with nothing more said, but think also about *why* these actions and institutions are ideal, then this unrealistic institutional design can be a useful heuristic for identifying the more fundamental principles that underlie not only this choice of institutional design for an idealised world, but also comparative judgements for the real world. Gilabert notes that Sen's distinction between descriptive and valuational proximity to an ideal is relevant here.[98] A mixture of red and white wine may be *descriptively* closer to pure red wine than pure white wine is, but not valuationally (assuming red wine is better than white), since the principles that make red wine better would in fact be better instantiated by the pure white wine than by the mixture. Thus, though identifying an ideal society purely descriptively might not be useful for making comparative judgements of possible institutional designs, identifying it valuationally would involve identifying what *makes* it the best society. Knowing this is relevant to real-world comparative judgements, since the principles that make the ideal society ideal are the same that make some feasible alternative better than another.

Thus, I do not think that Sen's arguments establish that unrealistic theory is not interesting or important, either in institutional design or theory of principles.[99] With theory of principles this seems clearly to be the case as, except where we can all agree that some proposal is a moral improvement, we need principles in order to make

[98] Gilabert (2012) 42

[99] There are other arguments against this view of Sen's in Gilabert (2008) and (2012); Jubb (2012); Robeyns (2012); Simmons (2010) and Stemplowska (2008). Gilabert argues, for instance, that if we focus only on comparative judgements we miss the importance of dynamic duties and the transitional standpoint; Robeyns and Simmons argue something similar. Gilabert (2012) also argues that ideal theory can be of inspirational and motivational significance and Jubb argues that ideal theory is important, since we need it to tell us when nonideal theory is tragic.

comparative judgements. While it may not strictly be *necessary* to seek the *most* fundamental principles underlying these, it does seem that the sort of theory that identifies the correct moral principles is important and relevant to guiding real-world action in this way. On the other hand, Sen is right that unrealistic institutional-design is not necessary or sufficient for institutional-design on lower, more realistic, FCs. It does seem like there is a certain level of outlandishness on the FC-scale, above which this sort of theory ceases to be straightforwardly relevant to real-world action. However, even here, it seems that there could be a useful role for very unrealistic institutional design: that of a heuristic. There are certain other ways in which theorists have attempted to argue that unrealistic theory is not a worthwhile form of theory to pursue. I do not have space to address them all, but I think the above arguments are sufficient to show how unrealistic theory can be of action-guiding use.

Conclusion

The concept of feasibility is often used in political philosophy and practical policy debate in a far too simplistic and cavalier manner. Quite apart from any doubts about when and whether feasibility ought to constrain our political theorising, feasibility critiques fail to have the force they are often taken to have simply because without further specification, they do not manage to say anything determinate. Furthermore, once we do make the meaning of 'feasibility' precise, it is then no longer obvious when the precisification arrived at ought to constrain our political philosophy. It does seem that infeasibility on some sharpenings can affect the truth of an 'ought'-claim, or a moral theory. However, it is certainly not required of every 'ought'-claim that what it calls for be feasible on *all* possible sharpenings of the term. Thus, it is not enough simply to object to a moral or political theory on the grounds that what it demands is not feasible on some chosen FC. We need, in addition, to argue that the theory (or part of a theory) needs to be feasible on *that* way of making the term precise. It has been thought (though not put in this language) that the sort of political theory that we should be interested in must be feasible on realistic or restrictive sharpenings of the term. If we constrain our theory only with feasibility defined in an expansive way, the recommendations we come up with will not be useful for the real world. I have argued here that this is not the case: theory constrained only by unrealistic feasibility constraints is interesting and important, as a guide or standard (when it seeks to identify the principles that underlie our particular policy or action decisions), and in order to achieve the optimal balance between conservatism and demandingness.

What can we say then about one of our examples of a theory often criticised on grounds of feasibility, participatory democracy? Feasibility critiques of participatory democracy may be of two sorts, as I suggested in my introduction. They may be straightforward,

claiming that it just is not feasible. The quote from Mill in the introduction suggests such a critique. Alternatively, they may be combined with considerations of desirability. The conjunction of a participatory democratic system and the realisation of certain other values is not feasible. The complaint attributed to Oscar Wilde that 'the trouble with Socialism is that it would take up too many evenings' (which is equally a complaint about any participatory system) seems to be of this sort. The idea is that a participatory system is not feasible in conjunction with certain other things we value: listening to music, talking to friends and family, and so on and so forth. In both cases, though, there is a feasibility claim being made, though the subject is different (in the first it is participatory democracy, in the second it is the conjunction of participatory democracy and having enough time to engage in valuable pursuits). In either case, whether the feasibility claim is true is not simply an empirical matter. It also depends on what exactly we mean by 'feasibility'. On some sharpenings of the term, these two outcomes are not feasible. If we hold fixed current preferences and motivations it seems like participatory democracy is not feasible, let alone in a desirable way. However, on certain other sharpenings, it seems likely that both outcomes are feasible. If we allow everything to vary but the laws of physics, it seems like a participatory system that does not take up too much time is feasible. With technology, participation in decision making can be made a minimal time commitment and it does not seem that there is any obvious reason why participatory democracy is inconsistent with the laws of physics (for example). Thus, a feasibility critique of participatory democracy that is supposed to warrant its rejection as a theory entirely must say that the only sharpenings that make it feasible are ones that we should not be interested in.

I have argued that we should be interested in theory constrained by a wide range of FCs. What this will mean exactly for participatory democracy, though, will depend on what sort of theory it is taken to be. I distinguished institutional design from theory of

principles, but these two are often combined or mixed up together in political theory. Political theorists often develop institutional designs at the same time as identifying the deeper moral principles that underlie the choice of institutional design (and explain why it is desirable). The two tasks are not always distinguished. On the one hand participatory democracy could be taken to be the theory of principles saying that we have a moral duty to respect the equal political participation of all citizens (and perhaps *to* participate) as much as possible, or something like that. Rousseau's theory of freedom as bound up with participation could be thought to be a theory of this nature. It then seems like, if its correctness is to be constrained by any FC, it will not be an overly restrictive one. Of course, I have not argued that the most fundamental principles are not constrained by *any* feasibility facts, and so I cannot claim that democratic participation is not unfeasible in a way that prevents it from being something we are fundamentally committed to. However, it seems unlikely that it is. It does not seem like political participation is something so unrealistic that it could not be called for by the fundamental principles about what we should do and how society should be.

Alternatively the theory of participatory democracy could instead be taken to be an institutional design. Rousseau's theory appears to have included institutional design as well (indeed, it is perhaps primarily of this type); he seems clearly to be setting out the sorts of institutions he thinks should govern society (in certain contexts). I have argued that there is an important role for at least somewhat unrealistic institutional design, to provide a goal and to make sure we are aware what is more desirable than what we have now, in order to be able to achieve the best balance between conservatism and utopianism. It is obviously not clear exactly how unrealistic an institutional design has to be to fail to be of any worth in this way, but it seems like an action design that tells us that participatory democracy is desirable given a somewhat expansive FC may be important. If it is correct, we are then aware that when more restrictive FCs demand

something else, it is not ideal, it is a second best. We can then attempt to make calculations about when it is better to pursue participatory democracy by attempting to change the feasibility facts, or by doing in the short term what will be likely to bring about the unrealistic recommendation in the long term, and when it is better instead to focus on the more realistic short-term institutional design. Of course, if participatory democracy is offered as a design for what we should implement immediately, then it is probably wrong, since recommendations for immediate action need to be constrained by more restrictive FCs. However, so long as we are aware what the purpose of a theory recommending participatory democracy is, there is an important role it can play. Feasibility considerations do not straightforwardly make it either wrong or uninteresting.[100]

[100] The shape and content of this thesis owe a lot to the comments, criticisms and suggestions of: an audience at the 2014 Brave New World political theory conference in Manchester; the participants in the UCL MPhil Stud thesis preparation seminar; Fiona Leigh; Michael Martin; Han Van Wietmarschen; José Zalabardo; and especially, my supervisor, Jo Wolff.

References

Austin, J.L. (1979) 'Ifs and Cans' in J.O. Urmson and G.J. Warnock (eds.) *Philosophical Papers* Oxford: Oxford University Press (originally published in 1956 in *Proceedings of the Aristotelian Society*)

Bales, R. E. (1971) 'Act-utilitarianism: account of right-making characteristics or decision-making procedures?' *American Philosophical Quarterly* 8: 257–65

Bentham, Jeremy (1961) *An Introduction to the Principles of Morals and Legislation* Garden City: Doubleday. Originally published in 1789

Berelson, B.R.; Lazarsfeld, P.F. and McPhee, W.N. (1954) *Voting: A Study of Opinion Formation in A Presidential Campaign* Chicago: University of Chicago Press

Brennan, Geoffrey and Pettit, Philip (2005) 'The Feasibility Issue' in Frank Jackson and Michael Smith (eds.) *The Oxford Handbook of Contemporary Philosophy* Oxford: Oxford University Press

Brennan, Geoffrey and Southwood, Nicholas (2007) 'Feasibility in Action and Attitude' in T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonsson (eds.) *Hommage à Wlodek: Philosophical Papers Dedicated to Wlodek Rabinowicz* www.fil.lu.se/hommageawlodek

Brighouse, Harry (2004) *Justice* Cambridge, UK: Polity Press

Buchanan, Allen (2004) *Justice, Legitimacy and Self-Determination: Moral Foundations for International Law* Oxford: Oxford University Press

Cohen, G.A. (2003) 'Facts and Principles' *Philosophy and Public Affairs* 31: 211-45

---------- (2008) *Rescuing Justice and Equality* Cambridge, MA: Harvard University Press

---------- (2009) *Why Not Socialism?* Princeton NJ: Princeton University Press

Cowen, Tyler (2007) 'The Importance of Defining the Feasible Set' *Economics and Philosophy* 23: 1-14

Daniels, Norman (2014) 'Justice and Feasibility', talk at the UCL Colloquium in Legal and Social Philosophy, Faculty of Laws, 19th March 2014

Elster, Jon (1985) *Making Sense of Marx* Cambridge and Paris: Cambridge University Press and Maison des Sciences de l'Homme

Estlund, David (2011) 'What Good is it? Unrealistic Political Theory and the Value of Intellectual Work' *Analyse & Kritik* 33: 395-416

---------- (2014) 'Utopophobia' *Philosophy and Public Affairs* 42: 113-34

Farrelly, Colin (2007) 'Justice in Ideal Theory: A Refutation' *Political Studies* 55: 844-64

Frankena, William (1950) 'Obligation and Ability' in Max Black (ed.) *Philosophical Analysis: A Collection of Essays* Ithaca, NY: Cornell University Press

Gheaus, Anca (2013) 'The Feasibility Constraint on the Concept of Justice' *The Philosophical Quarterly* 63: 445-64

Gilabert, Pablo (2008) 'Global Justice and Poverty Relief in Nonideal Circumstances' *Social Theory and Practice* 34: 411-38

---------- (2011) 'Debate: Feasibility and Socialism' *The Journal of Political Philosophy* 19: 52-63

---------- (2012) 'Comparative Assessments of Justice, Political Feasibility, and Ideal Theory' *Ethical Theory and Moral Practice* 15: 39-56

Gilabert, Pablo and Lawford-Smith, Holly (2012) 'Political Feasibility: A Conceptual Exploration' *Political Studies* 60: 809-25

Goodin, Robert E. (1995) 'Political Ideals and Political Practice' *Journal of Political Science* 25: 37-56

---------- (2012) 'The Bioethics of Second-Best' in Joseph Millum and Ezekiel J. Emanuel (eds.) *Global Justice and Bioethics* New York City: Oxford University Press

Hamlin, Alan and Stemplowska, Zofia (2012) 'Theory, Ideal Theory and the Theory of Ideals' *Political Studies Review* 10: 48-62

Hawthorn, G. (1991) *Plausible Worlds* Cambridge: Cambridge University Press

Hooker, Brad (2009) 'The Demandingness Objection' in Chappell (ed.) *The Problem of Moral Demandingness* New York City: Palgrave Macmillan

Jensen, Mark (2009) 'The Limits of Practical Possibility' *The Journal of Political Philosophy* 17: 168-84

Jubb, Robert (2012) 'Tragedies of non-ideal theory' *European Journal of Political Theory* 11: 229-46

Landauer, Carl (1959) *European Socialism: A History of Ideas and Movements, From the Industrial Revolution to Hitler's Seizure of Power* Berkeley and Los Angeles, CA: University of California Press

Lawford-Smith, Holly (2013) 'Understanding Political Feasibility' *The Journal of Political Philosophy* 21: 243-59

Lipsey, Richard G. and Lancaster, Kelvin J. (1956) 'The General Theory of Second Best' *Review of Economic Studies* 24: 11-33

Mason, Andrew (2004) 'Just Constraints' *British Journal of Political Science* 34: 251-68

Mill, John Stuart (1861) *Considerations on Representative Government* in H. B. Acton (ed.) *Utilitarianism, On Liberty, Considerations on Representative Government* (1972) London: J.M. Dent & Sons Ltd.

---------- (1863) *Utilitarianism* edited with an introduction by Roger Crisp. New York: Oxford University Press, 1998

Miller, David (1984) *Anarchism* London: J.M. Dent & Sons Ltd.

---------- (2008) 'Political philosophy for Earthlings' in David Leopold and Marc Stears (eds.) *Political Theory: Methods and Approaches* Oxford: OUP

Parfit, Derek (1984) *Reasons and Persons* Oxford: Clarendon Press

---------- (1997) 'Equality and Priority' *Ratio* 10: 202-21

Phillips, Michael (1985) 'Reflections on the Transition from Ideal to Non-Ideal Theory' *Nous* 19: 551-70

Pogge, Thomas (2008) 'Cohen to the Rescue!' *Ratio* 21: 454-75

Räikkä, Juha (1998) 'The Feasibility Condition in Political Theory' *The Journal of Political Philosophy* 6: 27-40

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

---------- (1999a) *A Theory of Justice: Revised Edition* Cambridge, MA: Harvard University Press

---------- (1999b) *The Law of Peoples* Cambridge, MA: Harvard University Press

Robeyns, Ingrid (2008) 'Ideal Theory in Theory and Practice' *Social Theory and Practice* 34: 341-62

---------- (2012) 'Are transcendental theories of justice redundant?' *Journal of Economic Methodology* 19: 159-63

Roemer, John E. (2002) 'Equality of opportunity: A progress report' *Social Choice and Welfare* 19: 455-71

Rousseau, Jean-Jacques (1996 [1762]) *Du Contrat Social* Paris : Librairie Générale Française

Russell, Bertrand (1910) 'The Elements of Ethics' in his *Philosophical Essays* London: Longmans, Green and Co.

Sen, Amartya (1997) 'Maximization and the Act of Choice' *Econometrica* 65: 745-79

---------- (1999) *Development as Freedom* Oxford: Oxford University Press

---------- (2006) ‘What Do We Want from a Theory of Justice?’ *The Journal of Philosophy* 103: 215-38

---------- (2009) *The Idea of Justice* Cambridge, MA: Bellknap Press

Sidgwick, Henry (1981 [1907]) *The Methods of Ethics* Indianapolis: Hackett Publishing Company

Simmons, A. John (2010) ‘Ideal and Nonideal Theory’ *Philosophy and Public Affairs* 38: 5-36

Smart, J.J.C. and Williams, Bernard (1973) *Utilitarianism: For and Against* Cambridge: Cambridge University Press

Stears, Marc (2005) ‘The Vocation of Political Theory: Principles, Empirical Inquiry and the Politics of Opportunity’ *European Journal of Political Theory* 4: 325-50

Stemplowska, Zofia (2008) ‘What’s Ideal About Ideal Theory?’ *Social Theory and Practice* 34: 319-40

Swift, Adam (2008) ‘The Value of Philosophy in Nonideal Circumstances’ *Social Theory and Practice* 34: 363-87

Valentini, Laura (2012) ‘Ideal vs. Non-Ideal Theory: A Conceptual Map’ *Philosophy Compass* 7/9: 654-664

Wiens, David (forthcoming) ‘Political Ideals and the Feasibility Frontier’ *Economics and Philosophy*

Wolff, Jonathan (1996) ‘Anarchism and Scepticism’ in John T. Sanders and Jan Narveson (ed.) *For and Against the State* Lanham, MD: Rowman and Littlefield, (1996), pp. 99-118

---------- (2011) *Ethics and Public Policy* London: Routledge

Wright, Erik Olin (2006) ‘Compass Points: Towards A Socialist Alternative’ *New Left Review* 41: 93-124