Theory, Ideal Theory and the Theory of Ideals

Alan Hamlin & **Zofia Stemplowska**
University of Manchester University of Reading

Abstract
The ideal/non-ideal distinction has commanded considerable recent attention. Its prominence is in large part due to the fact that it offers a vocabulary in which to diagnose what many commentators perceive as a problem at the heart of political theory: its relative unwillingness to provide solutions to urgent problems facing people here and now; or (to re-state the problem) its relative unwillingness to suggest institutional designs or reforms for people as 'they are' rather than as they 'should be'. The primary aim of this paper is to offer an improved understanding of the territory to which the ideal/non-ideal distinction relates. The core argument is that the ideal/non-ideal distinction operates only within a sub-region of the territory occupied by normative political theory, and that it both misses important parts of what is at stake in normative theorizing and presents too sharp a contrast between ideal and non-ideal theory. The main implication is that ideal theory is often assigned functions that it cannot and should not aim to deliver, while non-ideal theory is conceived, wrongly, as either 'applied' ideal theory, or a problem of second-best optimization that is not appropriately normative. Using conceptual tools familiar from the microeconomic analysis of choice, the paper suggests how to re-conceptualise the territory of normative theory and how those interested in institutional design and reform can draw on the work of political philosophers who do not themselves take up such questions.

Introduction

What do theorists do when they theorize? Merely asking the question reveals that there are many different activities that each have some claim to the title of theorizing: the construction of a set of statements designed to explain a specified group of facts; the construction of an hypothesis aimed at capturing key aspects of a causal mechanism; the construction of a prediction from a combination of known facts and other ideas; the proof of a theorem from explicitly stated axioms: any form of abstract reasoning; and so on. Our concern here is not with the choice among alternative general conceptions of theory, but rather with a much more specific issue that presents itself primarily in the realm of normative theorizing, where the distinction between ideal theory and non-ideal theory has commanded considerable attention.

Our primary aim is to offer an improved understanding of the territory that that this distinction relates to, in part by re-describing that territory in terms of the aims of theorizing rather than the specific properties of particular theories, and in part by entertaining the possibility of a third category which we shall term the theory of ideals. In overview, we argue that the ideal/non-ideal distinction operates only within a sub-region of the territory occupied by normative political theory, and that it both misses important parts of what is at stake in normative theorizing and presents too sharp a contrast between ideal and non-ideal theory. We develop the role of the theory of ideals and argue that the territory associated with the idea/non-ideal distinction is better viewed in terms of a multidimensional continuum ranging over a number of variables. Even though we aim to show that the ideal/non-ideal distinction is not the best available means of structuring our thoughts about the component elements of normative political theorizing, we will take it as our point of departure.

The main body of this essay is arranged in four further sections. The next section discusses various formulations of the ideal/non-ideal distinction, partly in order to offer criticism of each, and partly in order to sketch an alternative view of a continuum of possibilities that better represents the multi-dimensional nature of the issues at stake. We then turn to the theory of ideals, where we attempt to distinguish between developing an account of the ideals and principles that ground normative political philosophy (the role of the theory of ideals), and the derivation of implications for the structure and operation of political society that depends, in part (but only in part) on those ideals (the role of ideal and non-ideal theory). With these elements in place, the penultimate section directly addresses the question of the relationship between ideal and non-ideal theory and, specifically the question of whether ideal theory is a pre-requisite to non-ideal theory. Our answer here will be that it is not, but that it can act as a useful constraint on the prescriptions of non-ideal theory. Finally, we offer some concluding remarks.

The Ideal/Non-Ideal Distinction

While there seems to be widespread acceptance of the idea that the distinction between ideal theory and non-ideal theory is both useful and appropriate, there is little agreement on how exactly the distinction is to be specified. We initially identify three broad

approaches to the specification of the ideal/non-ideal distinction, where these approaches reflect[1]:

(1) the distinction between full compliance and non-full compliance
(2) the distinction between idealization and abstraction
(3) the distinction between fact-sensitivity and fact-insensitivity

We will outline and briefly discuss each of these in turn but, more generally, we argue that the relationship between these various approaches is complex and unclear and that each of them is attempting to ground a categorical distinction in an area where it seems much more appropriate to speak in terms of continuous variables. Compliance, idealization, abstraction and fact-sensitivity are all matters of degree and all matters of 'appropriateness'. While theories can certainly be compared according to their degrees of assumed compliance, or their degrees of idealization, abstraction and fact-sensitivity, and the appropriateness of their particular stances on these issues, any sharp or useful distinction between ideal and non-ideal categories of theory seems unlikely at best. Some theoretical approaches may be 'more ideal' than others, and for any particular question we might expect there to be a range of approaches which differ in their degree of 'idealness', so that an additional issue arises as to the relative advantages and disadvantages of alternative degrees of 'idealness' and, perhaps, the optimal degree of 'idealness' for the question and intended purpose in hand. A conceptual map of the ideal/non-ideal distinction that emerges as a result might thus be better construed as a multi-dimensional continuum.

(1) The distinction/continuum between full and non-full compliance.[2] A theory assuming full compliance assumes that more or less *everyone* does more or less *everything* that the normative content of the theory demands of him or her with respect to some domain. Given the presence of at least two variables – the number of compliers and the extent of compliance by each person – non-full compliance can take a number of forms.

Once we are within this continuum we can draw further distinctions that track the reasons for assuming a given level of compliance. David Estlund, for example, offers two further distinctions:[3] between hopeful and hopeless theory, and between aspirational and concecive theory. A theory is hopeless when it holds individuals (or institutions) to standards about which there is good reason to believe that they will never be met even when it would not be impossible to meet them (that is, there is good reason to believe that full compliance will not apply). A theory is hopeful when it holds individuals (or institutions) to standards about which there is no good reason to believe that they will not be met (where full compliance is at least feasible). A theory is aspirational when it posits standards that are currently not met but could be met. A theory is concessive if it concedes facts about how people and institutions are likely to act and guides action on the basis of this concession. All hopeless theories are therefore

[1] The following paragraphs draw on Stemplowska (2008), for related discussion see the other papers collected in that special issue of *Social Theory and Practice*, Farrelly (2007), Mills (2005), Goodin (1995), Nagel (1991) and Mason (2004).
[2] Rawls (1999): 7-8 and 212. Phillips (1985): 553-6; Murphy (1998): 278-9.
[3] Estlund (2008).

either utopian (unfeasible) or aspirational: they concede nothing and this is why they remain hopeless. Hopeful theories could be aspirational or concessive (or a mixture of both). It appears, then, that the distinction between hopeful and hopeless theory concerns what (there is good reason to believe) is and is not feasible (or likely), while the distinction between aspirational and concessive theory concerns whether any adjustments are made to what a theory recommends for the sole purpose of increasing the likelihood of the theory's requirements being complied with: aspirational theory does not make such adjustments while concessive theory does.

This illustrates a more general ambiguity relating to whether non-compliance is taken to be a matter of the infeasibility of full compliance, or just a matter of recognizing the probability of compliance in any particular setting. Resolving this ambiguity will often matter since a number of normative theorists have been less troubled by the first move (incorporating facts about infeasibility) than the second (incorporating facts about probability).[4] This might be taken to suggest that perhaps there is a point on the compliance/non-compliance continuum that usefully divides theory into ideal and non-ideal (so that we are dealing with a distinction after all): theories that assume some blameworthy or non-excused non-compliance are non-ideal while theories that assume no such non-compliance are ideal.[5] We think, however, that it would be a mistake to confine ideal theory to theory that assumes no blameworthy or non-excused non-compliance. We return to this issue below, for now we simply want to signal that there are a number of different points at which the level of compliance that distinguishes between the ideal and the non-ideal could be set.

(2) The distinction/continuum between idealization and abstraction (or, at least, the absence of idealization).[6] This distinction is in one sense more general than that relating to compliance, since it applies to all aspects of theory rather than just to the issue of compliance by individuals, but it also cuts across the issue of compliance. In its most basic form, abstraction is understood to consist in ignoring or bracketing off some complexities of a given problem, but without assuming any falsehoods about them. So that abstraction is taken to be a form of simplification in order to focus on the most important aspects of the question in hand. Idealization, on the other hand, consists in making false assumptions about some significant aspect of the problem in hand (O'Neill 1996: 40-1). Thus, for example, recommending that people be held responsible for their choices on the basis of the assumption that they can all choose wisely would involve an idealization (since we are assuming a falsehood about a number of people), while recommending that people be held responsible for their choices because it often has positive incentive effects – whether or not they can choose wisely – would involve mere abstraction from the complex reality in which some can and some cannot choose wisely.

Capturing what was implicit in O'Neill's discussion of idealization, Mills sets up the contrast between idealization and its absence in starker terms. According to Mills, we engage in idealization when we build a model of some P, which is not descriptive of what P is like, but rather models, on the basis of assumptions that are *significantly* false,

[4] Estlund (2008), Valentini (2009)
[5] Simmons (2010) esp. 8-9, 17 n.16,
[6] Farrelly (2007): 844-64, 848, O'Neill (1988): 55-69, O'Neill (1996): 38-44, Mills (2005): 165-84.

what an ideal P should be like (where 'should' can refer to a host of values: moral, prudential, etc.) (Mills 2005: 167-8). Ideal theory, according to this understanding of idealization, stipulates *significantly* false attributes to individuals and/or groups and their interactions, while non-ideal theory avoids doing this; but note that on this account, it is not just the inaccuracy of the assumption that matters, or its significance, but that this inaccuracy results from the notion of an 'ideal P'.

To illustrate the significance of this point imagine that we are concerned to model the motivation of agents, and we recognise that in the 'real world' there is considerable heterogeneity of motivation. Recognising the relevant degree of heterogeneity may make our model too unwieldy to be useful, so we consider adopting an assumption which limits the heterogeneity within the model. This is clearly a false assumption, but is it an idealization or an abstraction? One might answer that it is just an abstraction if we can defend the view that the assumption does not stem from any conception of what motivation individuals *should* adopt, or how much heterogeneity there *should* be, but the fact that such additional qualifications must be made illustrates our point that there is no straightforward distinction between idealisation and abstraction.

(3) The distinction /continuum between fact-sensitivity and fact-insensitivity. A theory is more fact-sensitive the more facts it recognizes and incorporates as elements of the model or as constraints on the model. Ideal theory is more fact-insensitive than non-ideal theory, and a criticism of ideal theory is that it is inappropriately fact-insensitive.[7] Of course, much depends how we understand 'facts' in this context. We might distinguish between contingent facts and necessary facts and suggest that insensitivity to at least some contingent facts may be what we previously termed 'abstraction', while insensitivity to necessary facts is 'inappropriate'. This merely shifts the argument on to the question of contingency/necessity, and the significance or appropriateness of particular facts, but this is sufficient to demonstrate that while the fact-sensitivity/fact-insensitivity basis for identifying the distinction between ideal and non-ideal theory is at least somewhat related to the idealization/abstraction basis, they are unlikely to be fully equivalent. We will pick up the issue of the appropriate idea of 'facts' when we discuss feasibility below.

There is a further distinction in the literature that is both of some importance and relates to the ideal/non-ideal distinction (or continuum, as we prefer to put it): the distinction between transcendental theory and comparative theory(Sen 2006).[8] A transcendental

[7] Farrelly (2007). For more general discussion see Cohen (2003).

[8] In recent work, Sen focuses on 'transcendental institutionalism' rather than simply transcendentalism (Sen 2009) A theory is transcendental if it focuses on indentifying 'perfect justice, rather than on relative comparisons of justice and injustice' (p5-6); it is institutional if it 'concentrates primarily on getting the institutions right, and it is not focused on the actual societies that would ultimately emerge' (p6). Sen admits, however, that transcendentalism and institutionalism need not go together (p6). We note an ambiguity in the understanding of institutionalism adopted by Sen. At times; institutionalism is understood simply as a concern with institutions and rules rather than 'non-institutional features, such as actual behaviours of people and their social interactions' (6p). But Sen also emphasizes that non-institutional theories are focused on actual realizations (p9) and so it appears that actual realizations are not the focus of institutional theories. However, if by 'being focused', Sen means 'regulate', then institutional

theory of justice, according to Sen, focuses on identifying perfectly just social arrangements, while a comparative theory concentrates on ranking alternative social arrangements. Put bluntly, the transcendental approach specifies the absolutely right or best case, while the comparative approach compares any two cases. Sen argues, correctly in our view, that there is no reason to suppose that either of these approaches subsumes or entails the other, but this is not our main interest in Sen's distinction.

At first glance it might be tempting to offer transcendental theory as an understanding of ideal theory, with comparative theory playing the role of non-ideal theory.[9] Such a view, for example, is adopted by Ingrid Robeyns when she argues that ideal theory is concerned with working out the principles of a perfectly just society[10], while '[o]ne important part of non-ideal theory is the development of principles for comparisons of justice in different social states (Robeyns 2008: 348). But we would argue that this temptation should be resisted. Sen's distinction between the transcendental and the comparative seems to us to pick out a rather different dimension of the theory enterprise and one that cross-cuts the ideal / non-ideal dimension.

We suggest that there is an ambiguity in Sen's usage that corresponds to the ideal/non-ideal distinction, so that Sen's discussion applies both within the set of ideal theories and within the set of non-ideal theories. Consider the transcendental case. As already indicated, the transcendental approach is characterised by its focus on the right or best. According to Sen, because the approach 'tries only to identify social characteristics that cannot be transcended in terms of justice, ...its focus is thus not on comparing feasible societies' (Sen, 2009: 6). But there is nothing in Sen's discussion that necessitates the interpretation of 'right' or 'best' or 'the most just' in terms that would confine them to the ideal theory end of the ideal/non-ideal continuum. Do we understand the transcendental approach to be limited to the 'ideal right or best', or the 'non-ideal right or best'? This question may be put for any of the interpretations if the ideal/non-ideal distinction discussed above. Whether we view compliance, idealization, abstraction, fact-sensitivity or any combination of these features as crucial to the ideal/non-ideal distinction, it is certainly possible (whatever Sen's intentions in the matter) to take a transcendental but non-ideal approach to the question of justice, by focusing all of our attention of the specification of the social arrangements that would represent maximal justice under the specified conditions. And Sen's contrast between the transcendental and the comparative would still be of significance under those specified conditions.

One reason why the transcendental/comparative distinction may initially appear to be related to the ideal/non-ideal distinction is that a reading of Sen suggests that the comparative approach is more suited to address questions of reform; that is policies,

---

theories may well be concerned with actual realizations, even when they aim to regulate only rules and institutions. If, on the other hand, by 'being focused', Sen means 'are concerned with', it is unclear whether there are any institutional theories, since even Rawlsians are concerned with non-institutional features of the societies they theorize about.

[9] Sen himself explicitly rejects such an interpretation (2009: p90) but it is notable that he adopts a very narrow understanding of the distinction between ideal and non-ideal theory, interpreting it as distinguishing between theories that assume compliance with reasonable behaviour and theories that do not.

[10] Robeyns (2008).

interventions or institutional modifications that relate to the possible reduction of injustice in the world as we know it without promising the delivery of any transcendentally just world. And of course this is plausible, but it depends on more than the distinction between the transcendental and the comparative, it also requires that the comparative approach be focused 'locally', both in terms of identifying policies, interventions or institutional reforms that are themselves feasible in some practical sense, and in terms of taking the 'world as we know it' as the basis for comparison. And it is precisely in these additional requirements of 'localness' that we ensure the relatively non-ideal flavour of the comparative method. It would be equally 'comparative' to address the relative justice of two or more hypothetical societies none of which approximated the world as we know it and where the comparison was independent of any notion of the feasibility of implementing reforms.[11]

We agree with Sen that the comparative approach is (generally) a necessary ingredient in any approach to the pressing problems of injustice, but it is by no means sufficient. There is also a requirement that the comparative method be applied at the relatively non-ideal level, which must be specified independently of the comparative focus of the theory.

The Theory of Ideals

Having suggested that the ideal/non-ideal distinction might be better construed as a multi-dimensional continuum, we now wish to introduce a rather different distinction, between, on the one hand, the theory of ideals and, on the other hand, that continuum of ideal and non-ideal theory.

Here the point at the heart of the distinction is the intended purpose of the theorizing. In the theory of ideals the purpose is to identify, elucidate and clarify the nature of an ideal or ideals (we will call this purpose 'specifying ideals'), whereas theorizing within the continuum of ideal to non-ideal theory is concerned with the identification of social arrangements that will promote, instantiate, honour or otherwise deliver on the relevant ideals (we will call this purpose 'institutional design').[12] One reason to think that institutional design is the aim of both ideal and non-ideal theory is that the debate over the degree of idealness that is appropriate is couched in terms of worries about impracticability and worries about short-term practicability over longer-term viability.

Now of course, there may well be some overlap or interplay between specifying ideals and institutional design. Serving either of the purposes may sometimes require making some set of assumptions about the extent of compliance, idealization, abstraction and fact-sensitivity. This is because one way of trying to specify an ideal is to ask what institutions and social arrangements it recommends, so as to be able to reflect on whether those institutional arrangements satisfy the intuition that lies behind the original

[11] Sen concedes as much: 2009:62

[12] There is, of course, the further question of what counts as a 'social arrangement' in this context. Specifically, what level of detail should be expected in the specification of a social arrangement or institution? Would Rawls's first principle of justice qualify as the identification of social arrangements? We lean towards the view that it would not; that Rawls is specifying the nature of the value of liberty and its priority so that this aspect of Rawls might be seen as part of the theory of ideals, rather than ideal theory. But nothing in our argument depends upon this.

specification of the ideal. But nevertheless, in this case, since the intention is to specify the ideal, we would categorise the theorizing as part of the theory of ideals.

To be a little more precise, we would suggest that there are essentially two component elements to the theory of ideals, one devoted to the identification and explication of individual ideals or principles (equality, liberty, etc.), the other devoted to the issues arising from the multiplicity of ideals or principles (issues of commensurability, priority and trade-off).[13] These aspects of the theory of ideals and their relationship with the issue of institutional design and, therefore the continuum between ideal and non-ideal theory can be illustrated diagrammatically.

Figure 1 represents the generalized problem of pluralist, consequentialist optimization in a manner that will be entirely familiar from the economist's analysis of choice. In the current context we take each axis to identify a particular value[14], and for illustrative purposes only we label these as J (for justice) and W (for welfare). The indifference curves (I_1, I_2, I_3) identify the trade-offs across values and so indicate levels of all-things-considered desirability. The feasibility frontiers (F_1, F_2, F_3) identify the outer limits of alternative sets of combinations of J and W that might be taken to be achievable. We take it that this figure illustrates the situation discussed in the final sentences of Cohen (2003). Our intention is not to make substantive points about the nature of the trade-offs between values, or of the general nature of the tension between feasibility and desirability. We also do not wish to imply that we can ever actually draw all relevant axes, full indifference curves and feasibility frontiers. Rather we use this diagram to identify and contrast the different senses of ideal and non-ideal theory and what we have termed the 'theory of ideals'.

How does theory contribute to Figure 1? Of course, in one sense Figure 1 depicts a particular theoretical conception, the conception of constrained maximisation, but we wish to decompose this conception into its various parts, so as to inspect the various ways in which ideal/non-ideal theory and the theory of ideals operate and interact within this over-arching conception. The theorist might begin[15] by attempting to specify a value: which amounts to identifying an axis in our diagram. The exercise of identifying a value, let us say J, and clarify its meaning and nature, may make little or no reference to other values (except insofar as such mention is required to distinguish J from those other values), but will be concerned with the structure of the value in question. For example, some values may be such that they may be fully realised, at least in principle. For the sake of argument we will take it that J is such a value and that J* in Figure 1 represents the full achievement of value J. Other values may be defined in such a way

[13] Cf. Swift (2008), Robeyns (2008). Swift also distinguishes between what we call the theory of ideals and ideal theory (when referring to the former he simply calls it philosophy). Also, like us, he observes that philosophy (in the context of normative theorising) has two tasks: it offers 'formal or conceptual analysis...[of] the various values at state, how they relate to one another, and so on...[and] substantive or evaluative judgements about the relative importance or value of the different values at stake.' (p369).

[14] We restrict attention to the case of two values so as to be able to use the familiar diagram, but all our points carry over straightforwardly to cases with more than two values.

[15] We make no claim regarding the logical or temporal ordering of the various theoretical elements that we identify, the sequence we adopt is purely for presentational convenience.

that we can imagine continuous and indefinite increases in that value. For the sake of argument, we will take it that value W is such a value, so that more W is always more desirable than less, ceteris paribus. In either case, the theorist will tackle the question of the appropriate measurement of the value, either in the interval up to its full realization, or over the full range. This aspect of the theorists work seems clearly to fit within the 'theory of ideals'[16].

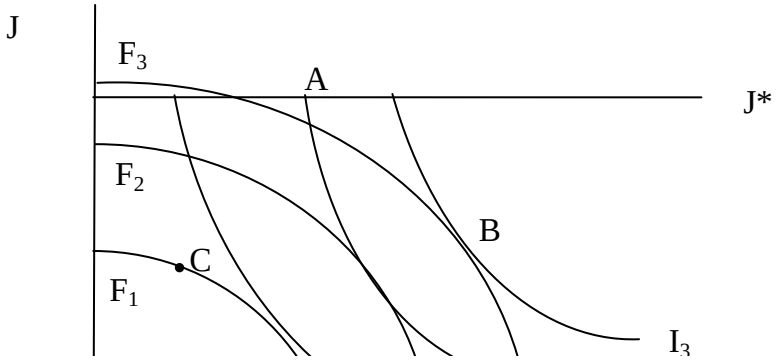

Figure 1 Generalized consequentialist optimization

Once we have a set of values specified in at least some detail and so have the axes and scales of our diagram, a next step might be to consider the interaction among values, including the nature and shape of any relevant indifference curves. If all relevant values were of the type that admit of full realization, we could think of the point at which these levels of full realization are mutually achieved (i.e. the intersection of the line J* with other equivalent lines) as a 'bliss point' or utopia; the all-things-considered ideal combination of values. But there is no general reason to suppose that such a point exists even in principle. If at least one relevant value is unbounded, in the way that we have assumed to be true of W in Figure 1, there will be no utopia or bliss point: movements to the right along the line at J* always increase overall value (as from I_1 to I_2 to I_3).

Given the indifference curves as shown, it is also the case that for any point on the line J*, such as A, where value J is fully achieved, we will be able to identify other points, such as B, that lie below J* but are nevertheless on a higher indifference curve, so that B is all-things-considered better than A, despite the fact that value J is fully realised at A but less than fully realised at B. Of course, indifference curves might not be as sketched. If one value is lexically prior to others there will be no indifference relationship that can be represented by a set of indifference curves. Such lexical priority over the full range of possibilities is surely extremely implausible, although it might be more plausible over some more local range. In any case, the theoretical discussion of the existence and shape of these indifference curves seems to fall naturally within the scope of what we have termed the theory of ideals.[17]

[16] We would point to works such as Cohen (2008) and Broome (1991) as excellent examples of this aspect of the theory of ideals.

[17] The shape of the indifference curves assumed in Figure 1 is familiar from standard economic models of consumer choice. The curvature shown is consistent with the idea of the diminishing marginal rate of substitution between the two values. That is, the rate at which the values are traded off against each other while holding all-things-considered value constant varies with the

While these aspects of the theory of ideals may be taken by some to represent 'ideal theory' in its most extreme form, since it takes no account of feasibility at all, we would argue that this is a form of category error, since we are not here engaged with the issue of institutional design at all. There is simply no reason for the theory of ideals to take account of issues of feasibility, since the inquiry is into the nature and structure of the normative criteria to be employed, just as in the economic analysis of consumer choice suggested by Figure 1, there is no reason to account for feasibility issues when contemplating the theory of utility functions and specifying their properties.

An objection might be pressed against this view. According to the objection, any specification of trade-offs between values must assume a fairly detailed specification of the scenarios in which the trade-offs between the values in question are to be judged. This is because considering trade-offs between values is only possible when we know what we are *really* giving up and gaining. Thus, we cannot compare an increase in equality against a decrease in privacy as such, we must instead compare a more equitable distribution of income (or some other good) against decreases in privacy that are meaningful to us: we must know what exactly it would mean to have the details of our salaries, expenses, or family dynamics accessible to the police; we must know how the police would use such data. In essence, all judgements of trade-offs are at bottom judgements over the desirability of concrete scenarios and any specification of concrete scenarios must assume particular feasibility constraints. We agree that thinking through concrete scenarios (actual and hypothetical) can be helpful, and might even be essential, to clarify what it is about a given value that is of value to us. But we disagree with the implied suggestion that interpreting the nature and structure of values (including trade-offs between them) must inevitably be done with a particular (more or less ideal) feasibility constraint in mind. On the contrary, we can only pursue the general inquiry into the nature and structure of values successfully if we are *not* tied to any particular feasibility constraint and are free to construct and compare hypothetical scenarios without reference to their feasibility or practicality. Assuming any particular feasibility constraint would give us only a very partial glimpse at our values; fuller inquiry precludes us from making such an assumption[18].

In parallel with the discussion of the various aspects of the theory of ideals, Figure 1 also invites theoretical discussion of the feasibility frontier, and it is here that we meet the continuum between ideal and non-ideal theory. Suppose that we find ourselves at a point such as C in Figure 1, how should we construct the relevant feasibility frontier? At one end of the range of possible approaches that we might adopt is assuming that C lies *on* the relevant frontier, as indicated by F_1. Such an assumption might be based on an

---

relative levels of the two values. Figure 1 assumes, in line with Rawls, that if society enjoys high welfare levels and relatively low levels of justice (equality), we might be willing to trade significant amounts of welfare in exchange for even a small increase in justice (and vice versa). Nothing crucial depends on the degree of curvature, and the argument holds when the marginal rate of substitution is constant, i.e. when indifference curves are straight lines. We acknowledge the possibility of lexical ordering of some values, and other criticisms of the formulation of Figure 1, later in the text.

[18] See Mason (2004) for an argument as to why justice in particular is not constrained by feasibility.

argument that is reminiscent of the economist's claim that 'there ain't no such thing as a free lunch'. If C were not on the feasibility frontier, it would necessarily be possible to increase both J and W simultaneously. Since such a move would be unambiguously good (a free lunch) it is difficult to see why the relevant actions had not been taken. An explanation might point to frictions or costs in the system, but if these costs are real costs (i.e. costs in terms of at least one of the values under consideration – W and J in this case) then this is just another way of saying that the actions to increase both J and W are not really feasible after all, since any attempt to act would incur costs that would lead to a reduction in either J or W or both. Of course, if the frictions are not real costs in this sense, the relevant actions are feasible, but then we are left with the original puzzle as to why they have not been taken.

We should be clear that we do not support or defend logic of this kind, we simply recognize it as identifying one extreme of the debate on the question of feasibility – the extreme that is most restrictive in setting the boundaries of feasibility or, put alternatively, the extreme that is most optimistic about the status quo: an optimism that is almost Panglossian, but not quite. Just because C (the status quo) is on the feasibility frontier, does not imply that it is optimal or the best of all possible worlds. Optimality is a matter of both feasibility and desirability – and a casual glance at Figure 1 suggests that C is not optimal within the feasibility frontier identified by F_1 and given the indifference curves as drawn. Movement around the feasibility frontier may improve all-things-considered desirability, even if it entails a reduction in one value (J, in this case).

Note that this almost-Panglossian approach to feasibility takes very seriously the limitations that may be imposed by individual character and by institutions[19]. Even if it is possible in some technical sense to imagine changing these aspects of society, the form of argument employed would suggest that such changes are typically costly and that any changes where the overall benefits exceed the overall costs might be expected to have been effected. This does not imply that there will be no change in the future, since the costs and benefits of various actions or institutional changes may change over time, but it may be held to provide at least some reason to think that the status quo is on the feasibility frontier given our current understanding of the costs and benefits of change. In this way this most restrictive feasibility frontier emphasizing all those factors that constrain choice in the here and now might be termed a short-run feasibility frontier.

At the other extreme of the feasibility debate we might hold the view that the only constraints on the achievement of J and W are those that are imposed by the (true) laws of science. In this case all that matters is what might be termed technical feasibility, and matters of apparent cost are deemed to be irrelevant (perhaps on the grounds that technology or other improvements in our understanding will, ultimately, show all such costs to be illusory). Such an account of feasibility would offer the most expansive account of the feasibility frontier (as might be depicted by F_3 in Figure 1) which might also be argued to correspond to the 'possible worlds' conception of feasibility. Here the

[19] Taking 'institutions' in the widest sense; to include laws, norms and other forms of social structures and relations.

status quo plays no significant role and, in particular, is not seen as the point from which changes must be costed. If an alternative social arrangement or an alternative account of the motivation of individuals is possible in the purely technical sense, then it is included in the relevant feasible set.

However, there is some vagueness about the meaning of 'technically possible' when considering issues such as individual motivation or institutional arrangements (as contrasted, say, with the 'technical possibility' of a perpetual motion machine). What are the limits of technical possibility in these domains? We might be able to imagine individuals who are motivated in some particular way, or social arrangements of a particular type, but we might still not recognise them as possible for 'us'. This can be so in two senses: (1) it might not be technically possible for us given path dependence and our history to date; (2) it might not be technically possible for us since it would require us to change into fundamentally different creatures. The tension between the imaginable and the truly reachable (as well as the tension between the imaginable for someone and the imaginable for us) lies at the heart of the issue on this construal of feasibility[20].

One aspect of Cohen's (2008) critique of the Rawlsian account of justice seems to hinge on the exact specification of what may be taken to be feasible in ideal theory. A simplified statement of the Rawlsian position, as summarised in the difference principle, identifies two relevant aspects of value, equality and the well-being of the worst-off group, and argues that full equity can be sacrificed if such sacrifice leads to an improvement in the well-being of the worst-off group. This may seem initially to be an argument in the theory of ideals, rather than in ideal theory but, for Cohen at least, the argument seems to hinge on the necessity of any possible trade-off between equality and the well-being of the worst-off group, rather than its desirability, and this in turn hinges on the issue of feasibility. In short, Cohen argues that the barrier to full equality must lie in the motivational structure of individuals, it is only because of some version of an incentive based argument that it is plausible to suggest that there might be circumstances in which the well-being of the worst-off group can be advanced by allowing inequality. But incentive arguments depend upon the motivations of the agents, and Cohen points out that if agents fully internalise the ideal of equality, and fully comply with its demands, there can be no tension between their incentives and full equality. So, if such agents are deemed to be feasible, it must be the case that full equality and the maximum well-being of the worst-off group can be realised simultaneously.

In terms of our diagrammatic presentation, this type of argument might suggest that the extreme case of the 'possible worlds' approach to feasibility will yield ideal feasibility frontiers that are rectangular – as in figure 2 where the ultimate or ideal feasibility frontier is given by F_4 which is in turn simply dictated by the greatest level of W that is achievable, given the laws of science, W*.

Between these two extremes lies a continuum of possibilities - each of which is more expansive than the Panglossian, but less inclusive than the 'possible worlds' approach. And it is this continuum, we suggest, that reflects the range from non-ideal to ideal

[20] See, for example, Brennan and Pettit (2005), Cowen (2007).

theory. Some parts of this continuum reflect different positions on individual motivation, for example the degree of compliance that one might expect with moral norms; other parts reflect the properties of institutions or social arrangements, for example the extent to which particular institutional structures gives rise to perverse incentives.

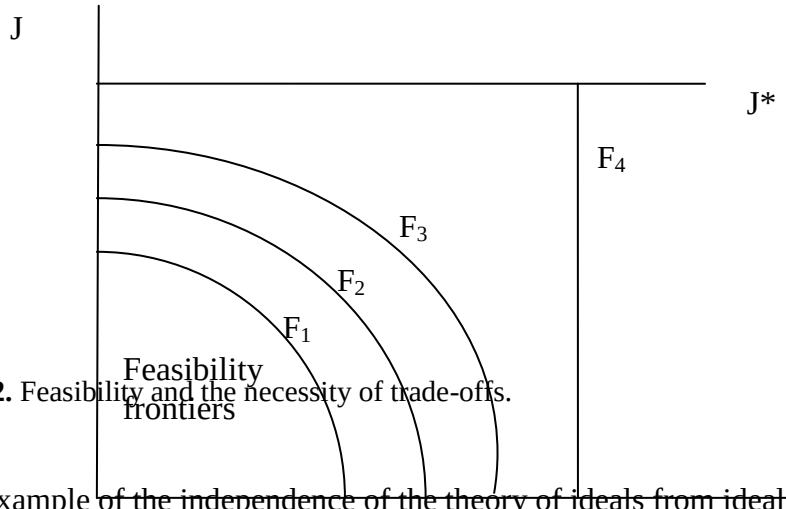

Figure 2. Feasibility and the necessity of trade-offs.

As an example of the independence of the theory of ideals from ideal theory, we might suggest that even if one accepted the feasibility argument just sketched in figure 2, so that, in the context of a fully ideal theory there is no necessary trade off between the values J and W, still, as a matter of the theory of ideals one could inquire into the nature of the indifference curves that identify the in-principle trade-offs between J and W. Of course, whatever indifference curves one superimposes on figure 2, if the relevant feasibility frontier is as shown by F_4, the optimum will lie at the intersection of J* and W*, but that does not make the theory-of-ideals specification of the indifference curves either impossible or irrelevant, since they will still be crucial in thinking about all non-ideal cases.

Three objections might be pressed against this way of conceiving of the continuum from non-ideal to ideal theory. First, some might argue that one form of ideal theory, a form that is not captured by the continuum, takes us beyond what is technically possible and into technical impossibility that is not feasible even in the longest run (and no matter where we started from). This form of ideal theory asks us to ignore what is, or is not, technically and motivationally possible in order to theorize about what is right, on the grounds that 'ought need not imply can'. In other words, we are asked to theorize about creatures whose motivational set-up may be technically impossible for us. One might react to this objection in several ways, but we would accept the substantive point without conceding the formal point. As we have already indicated, we take theorizing of the sort outlined in this objection as part of the theory of ideals concerned, as we put it, with the specification of ideals; rather than as part of the continuum from non-ideal to ideal theory which, as we put it, is concerned with institutional design. It is entirely plausible that the theory of ideals operates quite independently of any idea of feasibility; this is just to repeat and develop the point made above that in testing out our ideals we must be free to consider the implications of those ideals in situations that are entirely

hypothetical and agreed to be infeasible. But this does not imply that we might usefully draw recommendations for institutional design directly from such thought experiments.[21]

The second objection is one that was already sketched above: it holds that while there is a continuum of levels of compliance, there is a distinction between ideal and non-ideal theory in that any theory assuming some blameworthy/unexcused non-compliance qualifies as non-ideal while ideal theory is confined to theorising under assumptions of the absence of blameworthy/unexcused non-compliance. Indeed, the debate between Rawls and Cohen reported above seems to strengthen this objection since both authors explicitly accept that ideal theory solutions assume the absence of blameworthy/unexcused noncompliance.

We think that making such an assumption of compliance can be helpful, depending on the question we want to ask, but we do not think that proponents of ideal theory should make it one of the defining features of the theory. For notice that one consequence of making the assumption is that the problem of institutional design largely disappears from ideal theory. Why? Because if (almost) everyone does (almost) everything required of them by the relevant normative theory, the role for institutions in structuring and regulating behaviour seems relatively unimportant. This would imply that the nature of the problem of institutional design in the case of ideal theory is radically different from the nature of the problem of institutional design in non-ideal settings, and that the role of institutions in ideal theory is confined to informing and coordinating the actions of compliant individuals. Moreover, it implies that when designing requirements and institutions, ideal theory would not be able to take into account many of the costs that people incur in bringing their conduct in line with what is required of them, since the approach assumes that they are already motivated to act as they should. It is unclear why taking such costs into account cannot be seen as a task for ideal theory.

The third objection calls into question not only the ideal/non-ideal continuum but the suggested relationship between that continuum and the theory of ideals. Figure 1 represents an optimizing solution to the problem of institutional arrangements. Some might worry that it cannot be profitably used to illustrate the role of the theory of ideals and the ideal/non-ideal continuum for deontic, non-optimizing theories. In response, let us note, first, that deontic theories do not deny the relevance of consequential considerations; they simply deny that consequential considerations exhaust the set of relevant considerations. Under any plausible deontic account there will be an important role for consequential considerations and our discussion will apply directly in that domain. Or, to put it differently, our earlier discussion will apply straightforwardly to the domain of permissible actions that, alongside obligatory and impermissible ones, are part of deontic theories. Furthermore, and this is our second point, when modeling the obligatory and the impermissible, one useful technique is to impose side-constraints (either positive or negative) within the sort of diagrammatic model we sketch.

[21] Of course, there is also the issue of the logical limit of our imagination that constrains even the most expansive theory of ideals by facing us with what Parfit refers to as ‘deep impossibility’.

Feasibility frontiers could be used to incorporate such information and this implies that the model can be used represent deontic solutions to the problem of institutional design.

Is Ideal Theory a Pre-Requisite for Non-Ideal Theory?

There are many questions that can be asked within the setting sketched in the last two sections, and some styles of question may be more 'ideal' than others. For example, in the context of figure 1, we might ask whether full justice is feasible. This amounts to asking whether any part of the line J* lies within the feasible set. Such an inquiry will involve one aspect of the theory of ideals (the specification of J*) and some detailed account of feasibility. As figure 1 is drawn, and given the specification of J*, unsurprisingly, our answer to this question will depend on our view on feasibility: if we take an expansive view such as F_3 it is clear that J* is achievable, but a more restrictive accounts of feasibility (such as F_1 or F_2) will yield a negative response

But there is a further sense in which this question might be considered as part of 'ideal theory', since it enquires about an issue that is not focussed on the practical problem of optimal institutional design or of identifying the best feasible social arrangement. Even if we take the expansive view of feasibility embodied in F_3, so that J* is feasible, J* is not all-things-considered desirable, and a more practical question might focus on identifying the institutional arrangements and policy actions that might realise point B (the all-things-considered optimal point given F_3).

At base, we may identify the most practical, least 'ideal' theorizing as that which focuses attention on improvements from the status quo,[22] whether these improvements are seen as movements around a feasibility frontier in a manner that can be justified by appeal to all-things-considered value, or movements outward toward a feasibility frontier that represent gains in terms of all relevant values.

Keeping this in mind, we can turn to the key question we want to focus on in this section: is ideal theory a pre-requisite for non-ideal theory? We ask this question explicitly since we believe that it captures much of what is at stake in the literature devoted to the distinction between ideal and non-ideal theory, with one defence of ideal theory being the claim that it is indeed a pre-requisite for non-ideal theory. It should come as no surprise that we argue against this view. While elements of the theory of ideals should be seen as pre-requisites for both ideal and non-ideal theory, we suggest that theory that sits at any point in the ideal/non-ideal continuum may proceed without preliminary investment in 'more-ideal' theory[23]. That is, more practical, or 'non-ideal' theorizing needs to take as an input some understanding of the relevant values and some 'local' understanding of the interaction between values and feasibility. So, non-ideal theory will need at least some elements of a theory of ideals (though not necessarily a fine-grained or complete theory of ideals), but will not require ideal theory in the sense of a theory that operates on the basis of a more expansive specification of the feasible set.

[22] Jonathan Wolff emphasizes that policy-oriented political philosophy must make theorizing from the status quo part of its methodology, Wolff (2007) 128, 132-4; and 'Introduction' in Wolff (2010)

[23] Of course, this does not deny that there may be benefits of considering more-ideal and less-ideal theory in tandem in at least many cases, see next paragraph.

One argument that might suggest that ideal or more-ideal theory might be a pre-requisite for non-ideal or less-ideal theory is the argument from path dependence. If we conceive of less-ideal theory as aimed at discussing short run policy and institutional reforms taking seriously the feasibility constraints that seem significant here and now, while conceiving of more-ideal theory as aimed at identifying the long-term policy and institutional reforms that may become relevant in the future as feasibility constraints relax; then it might seem that we could view more-ideal theory as identifying a destination which our short-term reforms should keep in view. This might then imply that certain short-term policies that might appear desirable on the basis of less-ideal theory should be avoided if they set out on a path that is inconsistent with the long-term, more-ideal recommendations. In this way the results of more-ideal theory would serve as a guide to less-ideal theory.

Somewhat paradoxically, this line of argument suggests the imposition of a further class of constraints on less-ideal theory. Not only is less-ideal theory more heavily constrained by issues of feasibility, but it must also be constrained by the requirement of consistency with the recommendations of more-ideal theory.

While we agree that issues of path dependence may arise in particular circumstances, we do not think that this supports the general conclusion of the dependence of less-ideal theory on more-ideal theory. We would offer several counter arguments. First, we would dispute the generality of the essentially temporal view that less-ideal theory relates to the short-run, while more-ideal theory relates to the long-run. While some feasibility concerns may be temporal in this way, such that feasibility constraints relax over time (whether as a result of the pure passage of time, or as the result of time-related phenomena such as the advance of scientific understanding) others may have the opposite tendency with feasibility issues becoming more restrictive over time (for example, issues that might relate to reducing stocks of non-renewable materials, or rising populations), and still others may have no significant temporal dimension. The defining difference between less-ideal and more-ideal theory is logical rather than temporal, and this fact reduces the relevance and generality of the argument from path dependence.

Secondly, we do not believe that, even in those cases where path dependence may be an issue, we can assume that we have sufficient knowledge of the future path of feasibility constraints to effectively constrain less-ideal theory and its policy recommendations in any very specific way. Indeed, if we knew that something would be feasible in the foreseeable future it is difficult to see why we could not incorporate that fact into our less-ideal theorizing. If the mere possibility of future feasibility is to be taken as the basis for informing and constraining less-ideal theorizing and policy making, then we must ask about the temporal trade-off in costs and benefits that this implies. If we are to give up relatively certain gains in the short-term for the uncertain promise of larger gains in the long run, we would need a detailed and balanced view of the trade-off. And while this makes the point that, in such cases, there needs to be a dialogue between less-ideal and more-ideal theorizing, this is a genuine dialogue with less-ideal and more-ideal theory entering on an equal footing, rather than any claim that more-ideal theory is a prerequisite for less-ideal theory. More generally, the appropriate response to the

concern for possible path dependency problems when considering less-ideal theory and the question of policy analysis is to include in the analysis the value of keeping options open, or the cost of irreversible decisions.[24]

But despite our view that ideal theory is not a prerequisite for non-ideal theory, we also argue that there is significant value in pursuing theory at a range of points in the ideal/non-ideal continuum, some of which might be termed ideal, while others are termed non-ideal. Ideal theory that focuses on the more global or long-term issues of the nature of ultimate feasibility and their implications for institutional design, provides a check against the possibility that the pursuit of local or short term improvement might prove an ineffective means of pursuing global value (again, a theory of ideals offers a distinct and essential input). And just as more ideal theory may provide an important check on more practical non-ideal theory, so might non-ideal theory inform more ideal theory, by picking out those aspects of the theoretical structure that are most significant in practical terms, so directing the attention of ideal theorists. In this way ideal and non-ideal theory may be seen as deeply complementary while neither has priority over the other.

Conclusions

We might summarize our arguments as follows:

1. The ideal / non-ideal distinction may be better understood in terms of a categorical distinction between the theory of ideals (concerned with the specification of ideals) and the theory of institutional design that ranges over a continuum from the 'almost Panglossian' conception of feasibility to a 'possible worlds' conception of feasibility.
2. The multidimensional continuum conception of the domain of institutional design explains why there is a proliferation of more-or-less unsuccessful definitions of the 'distinction' between ideal and non-ideal theory: each definition tends to focus on one (or a small number) of a possible set of relevant dimensions.
3. Distinguishing even the most 'ideal' theory of institutional design from theory of ideals ensures that theorists do not miss out on proper analysis of ideals/values simply because they are worried that, since (they mistakenly believe that) they are operating within the non-ideal/ideal continuum, they should be careful about where they place the feasibility constraint while clarifying ideals/values.
4. Non-ideal theory is not 'applied' ideal theory but a separate problem, informed by elements of the theory of ideals.
5. However, although 'non-ideal theory' is not applied ideal-theory, this does not mean that it is not grounded in ideals or that it sells out on these ideals. This charge can take two forms. (1) Non-ideal theory is normatively impoverished in its understanding of ideals. This charge is misplaced because non-ideal theory can and should draw on the work of theory of ideals (as the diagram illustrates). (2) Non-ideal theory is concessive: it tells people what will suit them rather than

[24] For a classic discussion of the value of keeping options open in the context of public decision making see Arrow and Lind (1970). For a specific discussion of the costs of irreversible decisions see Arrow and Fisher (1974).

what they ought to do. But the charge in effect calls into question the relevance of 2nd best solutions, which is more than a reasonable person should want to bite. Just as long as we also have ideal theory, there is no reason to panic about undue concessions.

6. The role of ideal theory (or more ideal theory) is primarily to check for consistency and important inconsistencies in our advocacy of institutional and policy reforms as we consider alternative specifications of what is feasible. This allows us to consider short-run versus long-run reform and to engage in discussion which allows of the possibility that local optimization may not yield global optimization. It is not (primarily) to tell us what to do here and now and it is also not (primarily) to offer clarification of ideals/values.

References

Arrow, Kenneth J., and Anthony C. Fisher. 1974. Environmental Preservation, Uncertainty, and Irreversibility. *The Quarterly Journal of Economics* 88 (2):312-319.

Arrow, Kenneth J., and Robert C. Lind. 1970. Uncertainty and the Evaluation of Public Investment Decisions. *The American Economic Review* 60 (3):364-378.

Brennan, G, and P Pettit. 2005. The Feasibility Issue. In *The Oxford Handbook of Contemporary Philosophy*, edited by F. Jackson and M. Smith. Oxford: Oxford University Press.

Broome, John. 1991. *Weighing Goods*. Oxford: Blackwell.

Cohen, G. A. 2003. Facts and Principles. *Philosophy & Public Affairs* 31 (3):211-245.

———. 2008. *Rescuing justice and equality*: Harvard University Press.

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

Estlund, D. 2008. *Democratic Authority: A Philosophical Framework*. Princeton: Princeton University Press.

Farrelly, C. 2007. Justice in Ideal Theory: A Refutation. *Political Studies* 55 (4):844-864.

Goodin, R. E. 1995. Political ideals and political practice. *British Journal of Political Science* 25 (1):37–56.

Mason, A. 2004. Just Constraints. *British Journal of Political Science* 34 (02):251-268.

Mills, C. W. 2005. "Ideal Theory" as Ideology. *Hypatia* 20 (2005):165-184.

Murphy, L. B. 1998. Institutions and the Demands of Justice. *Philosophy & Public Affairs* 27 (4):251-291.

Nagel, T. 1991. The Problem of Utopianism. In *Equality and Partiality*, edited by T. Nagel: Oxford University Press.

O'Neill, Onora. 1988. Abstraction, Idealization and Ideology in Ethics. In *Moral Philosophy and Contemporary Problems*, edited by J. D. G. Evans. Cambridge: Cambridge University Press.

———. 1996. *Towards Justice and Virtue*. Cambridge: Cambridge University Press.

Phillips, M. 1985. Reflections on the transition from ideal to non-ideal theory. *Noûs* 19:551-570.

Rawls, J. 1999. *A Theory of Justice (revised edition)*: Oxford University Press.

Robeyns, I. 2008. Ideal theory in theory and practice. *Social Theory and Practice* 34 (3).

Sen, A. 2006. What do we want from a theory of justice? *The Journal of Philosophy* 103 (5):215-238.

———. 2009. *The idea of justice*: Belknap Press.

Simmons, A. J. 2010. Ideal and Nonideal Theory. *Philosophy & Public Affairs* 38 (1):5-36.

Stemplowska, Z. 2008. What's ideal about ideal theory? *Social Theory and Practice* 34 (3):319-340.

Swift, A. 2008. The value of philosophy in nonideal circumstances. *Social Theory and Practice* 34:363-387.

Valentini, L. 2009. On the Apparent Paradox of Ideal Theory. *The Journal of Political Philosophy* 17 (3):332-355.

Wolff, J. 2007. Harm and hypocrisy: Have we got it wrong on drugs? *Public Policy Research* 14 (2):126-135.

———. 2010. *Political Philosophy, Ethics and Public Policy*: Manuscript.