Political Studies REVIEW

Political Studies Association

POLITICAL STUDIES REVIEW: 2012 VOL 10, 48–62

doi: 10.1111/j.1478-9302.2011.00244.x

Theory, Ideal Theory and the Theory of Ideals

Alan Hamlin
University of Manchester

Zofia Stemplowska
University of Warwick

The recent prominence of the ideal/non-ideal debate is largely due to the fact that it offers a vocabulary in which to diagnose what many see as a key problem of political theory: its relative unwillingness to provide solutions to urgent problems facing people here and now; or for people as they are rather than as they should be. The primary aim of this article is to offer an improved understanding of the territory that the ideal/non-ideal debate relates to.

Keywords: ideal theory; non-ideal theory; feasibility; institutional design

Our primary aim is to offer an improved understanding of the territory that the ideal/non-ideal debate relates to, in part by re-describing that territory in terms of the aims of theorising rather than the specific properties of particular theories, and in part by introducing a further 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 theorising and presents too sharp a contrast. We develop the role of the theory of ideals and argue that the territory associated with the ideal/non-ideal distinction is better viewed in terms of a multidimensional continuum ranging over a number of variables. This last point explains, in our view, the proliferation of attempts to distinguish between ideal and non-ideal theory: there is no single, categorical and useful distinction to be found and what we see is a multiplicity of different ideas and debates which sit within a multidimensional terrain.

The main body of this essay is arranged in four further sections. The second section discusses various formulations of the ideal/non-ideal distinction. In the third section, we then sketch the theory of ideals and broach the general question of feasibility. With these elements in place, the fourth section directly addresses the question of the relationship between ideal and non-ideal theory and, specifically, the question of whether ideal theory is a prerequisite for 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 a summary of the key steps in the argument.

The Ideal/Non-ideal Distinction

While there seems to be widespread acceptance that the distinction between ideal theory and non-ideal theory is both useful and appropriate, there is little agreement on how exactly to specify it.[1] We identify four broad approaches to this problem of specification, with the distinction concerning:

(1) full compliance and non-full compliance;
(2) idealisation and abstraction;

(3) fact-sensitivity and fact-insensitivity;
(4) perfect justice (or another value) and local improvement in justice (or another value).

We will discuss each 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 more appropriate to speak in terms of continuous variables. Compliance, idealisation, abstraction, fact-sensitivity and improved realisation of a value are all matters of degree and of 'appropriateness'. While theories can certainly be compared across these dimensions, any sharp categorical distinction between ideal and non-ideal theories that focuses on a single dimension seems implausible at best. The conceptual map of the ideal/non-ideal territory that emerges is more usefully construed as a multidimensional continuum. For any particular question we might expect there to be a range of approaches that differ in their degree of 'idealness' in each of several dimensions, so that an issue arises as to the relative advantages and disadvantages of alternative degrees and patterns of 'idealness' and, perhaps, the optimal degree of 'idealness' for that question.

The Distinction between Full and Non-full Compliance[2]

Roughly, a theory assuming full compliance assumes that almost *everyone* does almost *everything* that the normative content of that theory demands of them. Given the presence of at least two variables – the number of compliers and the extent of compliance by each – non-full compliance can take a number of forms, giving us a continuum of (non-)compliance.

Once we are within this continuum we can draw further distinctions that track the reasons for assuming a given level of compliance. In his book *Democratic Authority*, David Estlund (2008, pp. 263–70) offers two further distinctions. The first is between hopeful and hopeless theory: a theory is hopeless when it holds individuals (or institutions) to standards that there is good reason to believe will never be complied with, even when it would not be impossible to comply. A theory is hopeful when it holds individuals (or institutions) to standards where there is no good reason to believe that they will not be complied with. Estlund's second and cross-cutting distinction is between aspirational and concessive theory. This distinction tracks whether the recommendations of a theory are adjusted for the purpose of increasing the likelihood of compliance. A concessive theory concedes facts about how people and institutions are likely to act and posits standards for action on the basis of this concession. An aspirational theory makes no such concessions, positing standards that are currently not complied with on the grounds that they ought to be.

Non-ideal theory is often associated with rejection of full compliance, but Estlund's distinctions highlight an ambiguity in any such rejection: is full compliance rejected because it is seen as impossible, or because it is seen as not probable in any particular setting? Resolving this ambiguity will often matter since a number of normative theorists have been less troubled by rejecting full compliance on grounds of impossibility than by rejecting full compliance on grounds of probability (Estlund, 2008; 2010; 2011; Valentini, 2009). What is troubling about the probable non-compliance move is that it invites concessive adjustments to a theory. This worry might suggest that there is a point on the compliance/non-compliance continuum that usefully divides theory into ideal and non-ideal categories (so that we are dealing with a categorical distinction after all). Such a point, if it exists, might be taken to define 'formal full compliance': theories that assume full compliance in at least this formal sense are 'ideal'; theories that assume that compliance falls short of even formal full compliance are 'non-ideal' (Simmons, 2010, pp. 8–9, p. 17 n. 16).

We think, however, that it would be a mistake to identify ideal theory with at least formal full compliance. More precisely, we think that making the assumption of formal full 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 a, still less *the*, defining feature of ideal theory. For one consequence of making this assumption is that the problem of institutional design largely disappears from view. After all, 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. True, institutions can still play the role of solving informational or coordination problems but they are no longer called for to incentivise, discourage or otherwise regulate behaviour. This would imply – implausibly – that the nature of the problem of institutional design in ideal theory is necessarily radically different from the nature of the problem of institutional design in non-ideal theory. Moreover, it would mean that when designing institutions (and other social arrangements), ideal theory is unable to take into account many of the costs that people incur in bringing their conduct in line with what is required of them, since it already assumes that they are motivated to act as they should. It is unclear, however, why taking such costs into account could not be seen as a task for ideal theory. It is unclear, for example, why ideal theory should be defined so that it could not deal with the problems of socialisation, preference formation or moral education.[3]

The Distinction between Idealisation and Abstraction (or, Perhaps, the Absence of Idealisation)[4]

Abstraction is understood to consist in bracketing off some complexities of a given problem, without assuming any falsehoods about them. It is a form of simplification undertaken to focus on the most important aspects of the problem in hand. Idealisation, by contrast, consists in making false assumptions about some significant aspect of the problem (O'Neill, 1996, pp. 40–1).

But the distinction between idealisation and abstraction is murky in practice. For example, imagine that we are concerned to include a treatment of the motivation of agents in our theory, 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 that limits the heterogeneity within the theory. This is clearly a false assumption, but is it an idealisation or an abstraction? One might answer that it is an abstraction if there is no reason to think that the assumption changes what we conclude from our theory. But how can we know? For problems where simplification is necessary but where the relationship between elements of the problem is unclear, there is no straightforward, and perhaps no helpful, distinction between idealisation and abstraction. This is not to deny that the distinction may be useful in at least some cases and contexts, but rather to suggest that the distinction is insufficiently clear in complex cases to act as a basis for the ideal/non-ideal demarcation. In our view, the idealisation/abstraction issue is best understood in terms of another complex continuum (of simplification) rather than founding a categorical distinction between ideal and non-ideal theory.

The Distinction between Fact-Sensitivity and Fact-Insensitivity

A theory is more fact-sensitive the more facts it recognises and incorporates as elements of the model or as constraints on the model. This immediately suggests another continuum, rather than a categorical distinction, so that the prospect of grounding an ideal/non-ideal distinction on fact-sensitivity seems remote.[5] One possibility is to argue that ideal theory is theory that is *inappropriately* fact-sensitive (Farrelly, 2007). But this only pushes back the question of what counts as *inappropriat*e fact-sensitivity. We might instead try to distinguish between contingent facts and necessary facts and suggest that only ideal theory can be insensitive to necessary facts. But this categorisation would not place what is often considered the paradigmatic ideal theory, namely Rawls' theory of justice, on the correct side of the divide (given his focus on a *realistic* utopia). That said, the thought that ideal and non-ideal theory can be analysed and understood in terms of a fact-sensitivity continuum is relatively close to our own proposal and we will address the issue of how to understand 'facts' when we discuss feasibility below.

The Distinction between a Theory of Perfect Justice (or Another Value) and a Theory of Local Improvement in Justice (or Another Value)

This distinction – also referred to as the distinction between transcendental and comparative theory – has been developed by Amartya Sen although he does not map it on to the ideal/non-ideal distinction.[6] Nonetheless, the distinction has been adopted by others for this purpose so it is relevant to consider it here.

A transcendental theory of justice focuses on identifying perfectly just social arrangements,[7] while a comparative theory concentrates on ranking alternative social arrangements. Put bluntly, the transcendental approach specifies the best case, while the comparative approach compares any two cases. It might be tempting to offer transcendental theory as an understanding of ideal theory, with comparative theory playing the role of non-ideal theory. Such a view, for example, is adopted by Ingrid Robeyns (2008, p. 348) when she argues that ideal theory is concerned with working out the principles of a perfectly just society, while '[o]ne important part of non-ideal theory is the development of principles for comparisons of justice in different social states'. We would argue that this temptation should be resisted.

To see why, notice that there is an ambiguity in Sen's framework. Consider the transcendental case. 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, p. 6). But there is nothing in Sen's discussion that necessitates the interpretation of 'right' or 'best' or 'the most just' in terms of the unfeasible or distant or absolute 'right' or 'best', rather than the feasible 'right' or 'best'. It is certainly possible (whatever Sen's intentions in the matter) to take a transcendental approach to the question of justice and yet focus all of our attention on identifying the social arrangements that would represent maximal justice under some particular non-ideal conditions (a local maximum). We can, for example, ask what the most just arrangement is given a particular form of non-compliance. Sen's response might be that, given the assumption of non-compliance, we may identify the most just arrangement but we could not identify a fully just arrangement (since the latter would exclude non-compliance). While this response is valid in one sense, it does not explain how we should classify attempts to identify a local maximum in relation to the transcendental/comparative distinction. The possibility of investigating local maxima shows that Sen's comparative/transcendental distinction leaves out forms of theorising that we expect to be able to categorise as ideal or non-ideal theory (or both).[8]

One reason why the transcendental/comparative distinction may initially appear to capture the elusive ideal/non-ideal distinction is that the comparative approach seems more suited to questions of reform: the design of policies, interventions or institutional modifications that offer a reduction of injustice in the world as we know it rather than promising a transcendentally just world. And of course this is plausible, but it requires that the comparative approach is complemented by a 'local' focus, both in terms of identifying policies, interventions or institutional modifications that are 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 hypothetical societies, neither of which approximated the world as we know it and where the comparison was independent of any notion of the feasibility of implementing reforms.[9]

To clarify, we accept that the distinction between the transcendental and comparative approaches can itself be categorical. But we also suggest that almost all of the work in making this distinction track any ideal/non-ideal distinction is being done by the assumptions of localness and realism that are imported into the comparative approach; and both localness and realism are surely better conceived as matters of degree.

We conclude this section by noting that while issues of compliance, idealisation, fact-sensitivity and comparability (as well as other issues) are all relevant to the ideal/non-ideal status of a theory, none of the discussions sketched succeeds either in establishing that any one issue holds the key to ideal/non-ideal status, or that this status is best considered in terms of a categorical distinction. Rather, we suggest that the territory over which the ideal/non-ideal debate ranges is better viewed as a multidimensional continuum.[10] Definitions of the ideal/non-ideal distinction that focus on one, or a small number, of a possible set of relevant dimensions may obscure other dimensions and so distort normative theory.

The Theory of Ideals and the Question of Feasibility

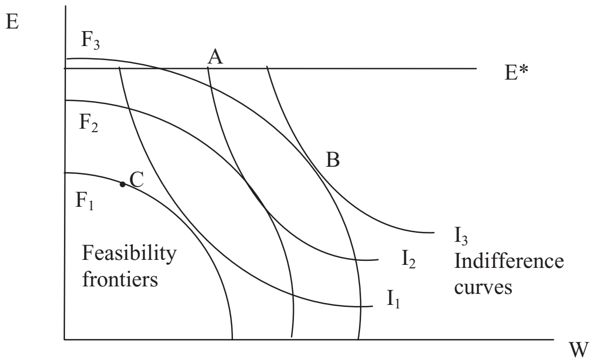
Having suggested that the ideal/non-ideal distinction is better construed as a multidimensional continuum, we now wish to introduce a rather different distinction, between, on the one hand, the theory of ideals and, on the other hand, that continuum of ideal and non-ideal theory.

At the heart of the distinction is the intended purpose of the theorising. In the theory of ideals the purpose is to identify, elucidate and clarify the nature of an ideal or ideals (we will call this 'specifying ideals'). More precisely, there are 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, trade-off, etc.). By contrast, the continuum of ideal/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 'institutional design' although we do not mean to imply that such design must concern coordinated human action; it can also concern practices engaged in by separate individuals).[11]

One reason to think that institutional design is the aim of both ideal and non-ideal theory (but not the theory of ideals) is that the debate over the degree of idealness that is appropriate is often couched in terms of worries about impracticability, and it is social arrangements rather than ideals that are subject to considerations of practicality. The mere fact that an ideal may not be perfectly realisable does not in itself serve to undermine it. That said, we are not suggesting that it is always easy to distinguish between 'ideals' and 'social arrangements'. Would Rawls' first principle of justice, say, qualify as a social arrangement or as an ideal? As it happens, we think it is the latter; Rawls is specifying the nature of the value of liberty and its priority, so that this aspect of Rawls' theory belongs to the theory of ideals, rather than ideal theory. But our point here is only that there is a genuine distinction between ideals (which one believes in, or not) and social arrangements (which one adopts, or not).

The relationship between the theory of ideals and the issue of institutional design can be illustrated diagrammatically. Figure 1 represents the generalised problem of pluralist, consequentialist optimisation 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

Figure 1: Generalised Consequentialist Optimisation



value,[12] and for illustrative purposes we label these as E (for equality) 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 value. The feasibility frontiers (F_1, F_2, F_3) identify the outer limits of alternative sets of combinations of E and W that might be taken to be achievable. We take it that this figure illustrates, more or less, the situation discussed in the final sentences of G. A. Cohen (2003) so that our general substantive question might be this: given that we value both equality and welfare, what social arrangements should we adopt (or justify)? Our intention, however, 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? The theorist might begin[13] by attempting to specify a value, which amounts to identifying an axis in our diagram. The exercise of identifying a value, say E, and clarifying its meaning and nature, may make little or no reference to other values (except in so far as such mention is required to distinguish E 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 E is such a value and that E★ in Figure 1 represents its full achievement. Other values may be defined in such a way that continuous and indefinite increases in that value are possible. 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, we tackle the question of the appropriate measurement of the value and this work seems clearly to fit within the 'theory of ideals'.[14]

Once we have specified a set of values and so have the axes and scales of our diagram, we might consider the interaction among values, including the nature and shape of any relevant indifference curves.[15] If all relevant values, like E, admit of full realisation, we could identify the point at which all values are simultaneously fully realised (i.e. the intersection of the line E★ with equivalent lines) as a 'bliss point' or all-things-considered ideal. But such a point may not exist, even in principle. If at least one relevant value is unbounded, as we have assumed to be true of W, there will be no bliss point: movements to the right along the line at E★ always increase overall value.

Given the indifference curves as shown, it is also the case that for any point on the line E★, such as A, where value E is fully achieved, we will be able to identify other points, such as B, that lie below E★ but are nevertheless on a higher indifference curve, so that B is all things considered better than A, despite the fact that value E 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 local range. In any case, the theoretical discussion of the existence and shape of these indifference curves falls naturally within the scope of the theory of ideals.

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.

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 welfare as such; we must instead compare a more equitable distribution of income (or some other good) against specified decreases in welfare that are meaningful to us. 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, in clarifying what it is about a given value that is of value to us. But we disagree that interpreting the nature and structure of values (including trade-offs between them) must inevitably be done with a particular 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. Assuming any particular feasibility constraint would give us only a very partial glimpse of our values; fuller inquiry precludes us from making such assumptions.

Indeed, notice that even if we accept, as most political theorists probably do, that the value of justice is constrained by what is feasible – so that a truly unfeasible requirement cannot be a requirement of justice (Miller, 2008) – it would still not follow that in specifying the ideal of justice we must not venture beyond what is feasible. This is because to understand the ideal of justice fully it may be important to ask what justice would require in the absence of the relevant feasibility constraint (Cohen, 2008, pp. 252–4; Mason, 2004).[16] It matters, that is, to our understanding of justice whether some requirement is not a requirement of justice *merely* because satisfying it is not feasible, or because it would not be required by justice anyway. For example, it may well not be feasible for all parents to give up their children happily. But we do not understand parental justice fully unless we ask whether justice would require this of parents if it became feasible.

But Figure 1 invites theoretical discussion of the feasibility frontier, even if selecting a specific frontier is not a part of the theory of ideals. It is here that we meet the continuum between ideal and non-ideal theory. Supposing 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 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 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 must be possible to increase both E 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 (costs, for example, that result in short-termism or short-sightedness), but if these costs are real costs (i.e. costs in terms of at least one of the values under consideration – E and W in this case) then this is just another way of saying that the actions to increase both E and W are not really feasible after all, since any attempt to act would incur costs leading to a reduction in E and/or W. 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 recognise 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 another way, the extreme that is most optimistic about the status quo: an optimism that is almost Panglossian, but not quite. 'Not quite', just because C (the status quo) is on the feasibility frontier, it does not imply that it is optimal or the best of all possible worlds. Optimality is a matter of both feasibility and desirability – and Figure 1 suggests that C is not optimal within the feasibility frontier F_1 given the indifference curves as drawn.

Note that this almost Panglossian approach to feasibility takes very seriously the limitations that may be imposed by individual character and social arrangements. Even if it is possible to imagine changing these aspects of society, this approach suggests that such changes are typically costly and 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 offers a 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 emphasising all those factors that constrain choice here and now might be termed a short-run feasibility frontier.

At the other extreme of the feasibility debate lies the view that the only constraints on the achievement of E and W are those imposed by the true laws of science. In this case all that matters is what might be termed 'technical feasibility', and apparent costs are deemed 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 offers the most expansive account of the feasibility frontier (as might be depicted by F_3) which might be thought to correspond to the 'possible worlds' conception of feasibility. Here the 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 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.[17]

To illustrate the importance of the above discussion, note that one of the key debates in recent political theory concerns precisely the expansiveness of the appropriate feasibility frontier. We have in mind here Cohen's (2008) incentive critique of Rawls. A simplified statement of the Rawlsian position, as summarised in the difference principle, identifies two values that are relevant to justice: equality and the well-being (measured in primary goods) of the worst-off group. Rawls argues that full equality can be sacrificed if this leads to an improvement in the well-being of the worst-off group. The trade-off between equality and well-being suggested by Rawls may seem to invite an examination of his theory of ideals since it is the latter that specifies how much of one value to trade for another. But Cohen's famous response has been to deny the necessity of any possible trade-off between equality and well-being, and this hinges on the issue of feasibility. In short, Cohen's objection to Rawls is that if we assume, following Rawls, that individuals are motivated to comply with justice, then the need to trade off equality and well-being disappears. It only arises in the first place because talented people demand incentive payments to become more productive. But people who are motivated to realise justice fully would not demand incentive payments but rather increase productivity without them. 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 effect, Cohen's approach to feasibility yields a single feasibility frontier that is rectangular. In terms of Figure 1, it would consist of the line E* and a new, vertical line at W* – the highest level of W that is technically achievable given the laws of science.

Cohen takes one position on feasibility while Rawlsians can take another. Still other positions are available along a continuum of possibilities, from the Panglossian to the 'possible worlds' approach. And it is this continuum, we suggest, that reflects the range from non-ideal to ideal theory. In this sense, Cohen adopts a maximally ideal stance.

Two important objections might be pressed at this point against our understanding of the terrain of normative theory. First, some might argue that one form of ideal theory not captured by our understanding of the ideal/non-ideal continuum takes us beyond what is technically possible and into technical impossibility: ignoring technical and motivational possibility in order to theorise about what is right, on the grounds that 'ought need not imply can'. In other words, we are asked to theorise about worlds subject to different laws of science.

One might react to impossibly ideal theory in several ways, but we would accept the substantive point without conceding the formal point. As we have already indicated, we take theorising without constraints of feasibility to be part of the theory of ideals, rather than part of the continuum from non-ideal to ideal theory concerned with institutional design: in testing out our ideals we must be free to consider the implications of those ideals in situations that are entirely hypothetical. But this does not imply that we might usefully draw recommendations for institutional design directly from such thought experiments.[18]

The second objection questions the framework within which we have presented the relationship between the ideal/non-ideal continuum and the theory of ideals. Our discussion and Figure 1 operate within a teleological and optimising approach. Some might worry that the ideal/non-ideal continuum and the theory of ideals framework is limited to such an approach and, in particular, that it may not apply to a deontic approach. In response, note first that deontic theories do not deny the relevance of teleological considerations; they simply deny that teleological considerations exhaust the set of relevant considerations. Any plausible deontic account will grant an important role to optimising considerations. In this way, our discussion will apply straightforwardly to the domain of permissible actions that, alongside obligatory and impermissible actions, form part of deontic theories. Furthermore, the distinction between the theory of ideals and the domain of institutional design, which ranges over the ideal/non-ideal continuum, is also helpful when considering the obligatory and the impermissible elements of a deontic theory. After all, it is common practice to conceptualise deontic rules as forms of (positive and negative) constraints, and the interaction of constraints with ideals is precisely the basis of our analytic apparatus.

Is Ideal Theory a Prerequisite for Non-ideal Theory?

Various normative questions can be framed within the setting sketched above, and some questions may be more ‘ideal’ than others. For example, we might ask whether full equality is achievable. Our theory of ideals has already delivered a partial answer by specifying equality as a value that is at least in principle capable of being fully realised (at E★), so that the rest of the answer depends on the precise specification of the feasible set. As Figure 1 is drawn, if we take an expansive view such as F_3, E★ is achievable, but a more restrictive account of feasibility (F_1 or F_2) will yield a negative response. But the feasibility of E★ does not settle the question of the all-things-considered desirability of E★. So our institutional design question must distinguish between identifying the feasible social arrangements that achieve full equality, and the feasible social arrangements that are best all things considered.

At base, we may identify the most practical, least ‘ideal’ theorising as that which focuses attention on improvements from the status quo,[19] whether these improvements are seen as movements around a feasibility frontier for the sake of all-things-considered value, or movements outward toward a feasibility frontier that represents gains in all relevant values.

Keeping this in mind, we can turn to the key question of this section: is ideal theory a prerequisite 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 ideal/non-ideal distinction, with one defence of ideal theory being that it is such a prerequisite (Simmons, 2010). It should be no surprise that we disagree. While elements of the theory of ideals should be seen as prerequisites for both ideal and non-ideal theory, theory that sits at any point on the ideal/non-ideal continuum may proceed without preliminary investment in ‘more-ideal’ theory. More practical, less-ideal theory needs to take as an input some account of the relevant values and of the interaction between values and feasibility, though not necessarily a fine-grained or complete theory of ideals, but does not require more-ideal theory in the sense of a theory that operates on the basis of a more expansive specification of the feasible set.

One argument which might suggest that more-ideal theory is a prerequisite for less-ideal theory is the argument from path dependence. If we conceive of less-ideal theory as aimed at identifying short-term reforms that take seriously the feasibility constraints that bind here and now, while conceiving of more-ideal theory as aimed at identifying long-term reforms that become relevant if feasibility constraints relax, then it might seem that we could view more-ideal theory as identifying a destination that our short-term reforms should keep in view. This might then imply that certain short-term reforms which appear desirable on the basis of less-ideal theory should be avoided if they set out on a path that is inconsistent with longer-term, more-ideal recommendations. In this way the results of more-ideal theory would serve as a guide to less-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 offer two counter-arguments. First, we 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 such that feasibility constraints relax over time (perhaps alongside the advance of scientific understanding), others may have the opposite tendency with feasibility issues becoming more restrictive over time (for example, due 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.

Second, 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 theorising.[20] If the mere possibility of future feasibility is to be taken as the basis for informing and constraining less-ideal theorising 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 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 theory, this is a genuine dialogue with each 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.[21]

Summary

We summarise our major points 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 which ranges over a continuum from the 'almost Panglossian' conception of feasibility to the 'possible worlds' one.

(2) The multidimensional continuum conception of the domain of institutional design explains the proliferation of more-or-less unsuccessful definitions of a categorical 'distinction' between ideal and non-ideal theory: each definition tends to focus on one (or a small number) of the set of relevant dimensions.

(3) Non-ideal theory is not 'applied' ideal theory but is simply the study of a different problem.

(4) 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 these ideals. This charge can take two forms. (a) 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 theory of ideals. Distinguishing ideal/non-ideal theory from the theory of ideals helps ensure that theorists do not miss out proper analysis of values because they mistakenly believe that they must stick to the feasibility constraint they adopt for institutional design even when clarifying the values at stake. (b) Non-ideal theory is concessive: it tells people what suits them rather than what they ought to do. This charge is misplaced since second-best solutions can still be challenging.

(5) A key role of ideal theory (or more-ideal theory) is to check for consistency in our advocacy of institutional and policy reforms as we consider alternative specifications of the feasible. This allows us to consider short-run versus long-run reform, local versus global optimisation, path dependency and related issues. It is not (primarily) to tell us what to do here and now, and it is not (primarily) to offer clarification of ideals/values.

(*Accepted*: 7 October 2011)

About the Authors

Alan Hamlin is Professor of Political Theory and Head of Politics at the University of Manchester. He has published widely in the areas of rational actor political theory, public choice, constitutional political economy and institutional design. His current work focuses on expressive political behaviour and the analysis of conservatism. Alan Hamlin, Politics, School of Social Sciences, University of Manchester, Manchester M13 9PL, UK; email: *alan.hamlin@manchester.ac.uk*

Zofia Stemplowska is Associate Professor of Political Theory at the University of Warwick. Her most recent publication is a co-edited book (with Carl Knight) entitled *Responsibility and Distributive Justice* (Oxford University Press, 2011). Zofia Stemplowska, Department of Politics and International Studies, Social Sciences Building, The University of Warwick, Coventry CV4 7AL, UK; email: *z.t.stemplowska@warwick.ac.uk*

Notes

We are grateful for comments to the participants of the 'Justice between Ideals and Reality' workshop, 'Justitia Amplificata', Frankfurt (January 2010), 'Ideal and Nonideal Theory' session, Central APA, Chicago (February 2010), 'Nonideal and Institutional Theory' workshop, CPSA, Montreal (June 2010), 'Democracy and Utopophobia conference', CONCEPT, Nottingham (December 2010) and the anonymous reviewers of *Political Studies Review*.

[1] The following paragraphs develop Stemplowska (2008).

[2] Murphy, 1998, pp. 278–9; Phillips, 1985, pp. 553–6; Rawls, 1999, pp. 7–8, p. 212; Sen, 2009, p. 90.

[3] After all, Rawls is considered an ideal theorist while tackling issues of preference formation, etc. See, for example, Part III of *A Theory of Justice* (Rawls, 1999).

[4] Farrelly, 2007, pp. 844–64, p. 848; Mills, 2005, pp. 165–84; O'Neill, 1988, pp. 55–69; 1996, pp. 38–44; Valentini, 2009, pp. 227–40.

[5] Fact-sensitivity/insensitivity may be thought to map on to abstraction/idealisation but, in our view, the relationship is less than straightforward. The problem is beyond the scope of this essay since in any case we do not see the abstraction/idealisation issue as a good basis for understanding ideal and non-ideal theory.

[6] Sen 2006; 2009, p. 90. See Note 2 above.

[7] In recent work, Sen (2009, pp. 5–6) focuses on 'transcendental institutionalism'. A theory is transcendental if it focuses on identifying 'perfect justice, rather than on relative comparisons of justice and injustice'; 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'. Sen admits that transcendentalism and institutionalism need not go together.

[8] For further discussion of the limits of the comparative approach if unaided by the transcendental, see Estlund (2011).

[9] Sen concedes as much (2009, p. 62).

[10] Estlund (2008) hints at the same possibility.

[11] Some might suggest that 'institutional design' is a narrow label since it rules out, for example, the radical anarchist who focuses on issues of individual behaviour and eschews 'institutions'. But we would argue that anarchism is a form of institutional design even if the institutions that are advocated are minimal or even what is suggested is their abolition. An alternative to 'institutional design' would be 'action guiding' but we prefer 'institutional design' precisely because of its focus on social and political arrangements. We are grateful to an anonymous *PSR* reviewer for making us clarify this point. More generally, see Robeyns, 2008; Swift, 2008. Swift also distinguishes between what we call 'the theory of ideals' and 'ideal theory' (he calls the former 'philosophy'). 'Philosophy' offers 'formal or conceptual analysis ... [of] the various values at stake, 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' (Swift, 2008, p. 369).

[12] We focus on two values to be able to use the familiar diagram, but all our points carry over straightforwardly to cases with more values.

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

[14] Cohen (2008) and Broome (1991) are excellent examples of such work.

[15] The shape of the indifference curves assumed in Figure 1 is familiar from economic models of consumer choice. The curvature shown is consistent with a diminishing marginal rate of substitution between the two values. That is, the rate at which the values are traded off against each other holding the level of all-things-considered value constant varies with the relative levels of the two values. Nothing crucial depends on the degree of curvature, and the argument holds when the marginal rate of substitution is constant and indifference curves are straight lines.

[16] David Miller's position is that a constraint of feasibility (of the specific type that he endorses) defines the boundaries of any attractive conception of justice, so that what is not feasible (in his sense) cannot, by definition, be a requirement of justice. More generally, therefore, he could claim that the theory of ideals should be capable of specifying values that incorporate a feasibility constraint as part of their definition. Indeed, some conceptions of justice assume specific feasibility constraints (e.g. agreement by actually existing reasonable people) and such conceptions already dismiss some feasibility frontiers as irrelevant. (They do not fix on a specific feasibility frontier, but narrow the range of relevant feasibility frontiers.) So a discussion of feasibility may be part of the theory of ideals since it helps us to specify the value of justice. We can accept this final point without accepting that this reduces the relevance of the distinction between the theory of ideals and the theory of institutional design.

[17] Brennan and Pettit, 2005; Cowen, 2007. Our concern here has been focused on the logic and structure of the issue of the relationship between the theory of ideals and the continuum of approaches to institutional design, and the role of the idea of feasibility in that logic. Clearly there is much more to be said about the relevant content of the idea of feasibility (and the related set of ideas about realism) in political theory. See, for example, Cohen, 2008; Galston, 2010; Miller, 2008; Philp, 2010; Stemplowska and Swift, forthcoming; Ypi, 2011, ch. 2.

[18] Note that even the most prominent advocate of fact-independent principles did not think so (Cohen, 2008).

[19] Wolff (2007) holds that theorising from the status quo is essential for policy-oriented theory.

[20] This is compatible with rejecting Sen's point that the comparative/less-ideal approach has no business knowing the end point in view.

[21] 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).

References

Arrow, K. J. and Fisher, A. C. (1974) 'Environmental Preservation, Uncertainty, and Irreversibility', *The Quarterly Journal of Economics*, 88 (2), 312–9.

Arrow, K. J. and Lind, R. C. (1970) 'Uncertainty and the Evaluation of Public Investment Decisions', *The American Economic Review*, 60 (3), 364–78.

Brennan, G. and Pettit, P. (2005) 'The Feasibility Issue', in F. Jackson and M. Smith (eds), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press, pp. 258–79.

Broome, J. (1991) *Weighing Goods*. Oxford: Blackwell.

Cohen, G. A. (2003) 'Facts and Principles', *Philosophy & Public Affairs*, 31 (3), 211–45.

Cohen, G. A. (2008) *Rescuing Justice and Equality*. Cambridge MA: 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 NJ: Princeton University Press.

Estlund, D. (2010) 'Human Nature and the Limits (if Any) of Political Philosophy'. Paper presented at Canadian Political Science Association, Montreal, 3 June.

Estlund, D. (2011) 'The Best and the Rest: Optimizing and Comparing in Theories of Justice', unpublished manuscript.

Farrelly, C. (2007) 'Justice in Ideal Theory: A Refutation', *Political Studies*, 55 (4), 844–64.

Galston, W. A. (2010) 'Realism in Political Theory', *European Journal of Political Theory*, 9 (4), 385–411.

Mason, A. (2004) 'Just Constraints', *British Journal of Political Science*, 34 (2), 251–68.

Miller, D. (2008) 'Political Philosophy for Earthlings', in D. Loeopold and M. Stears (eds), *Political Theory: Methods and Approaches*. Oxford: Oxford University Press, pp. 29–48.

Mills, C. W. (2005) ' "Ideal Theory" as Ideology', *Hypatia*, 20, 165–84.

Murphy, L. B. (1998) 'Institutions and the Demands of Justice', *Philosophy & Public Affairs*, 27 (4), 251–91.

O'Neill, O. (1988) 'Abstraction, Idealization and Ideology in Ethics', in J. D. G. Evans (ed.), *Moral Philosophy and Contemporary Problems*. Cambridge: Cambridge University Press, pp. 55–69.

O'Neill, O. (1996) *Towards Justice and Virtue*. Cambridge: Cambridge University Press.

Phillips, M. (1985) 'Reflections on the Transition from Ideal to Non-ideal Theory', *Noûs*, 19, 551–70.

Philp, M. (2010) 'What is to be Done? Political Theory and Political Realism', *European Journal of Political Theory*, 9 (4), 466–84.

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

Robeyns, I. (2008) 'Ideal Theory in Theory and Practice', *Social Theory and Practice*, 34 (3), 341–62.

Sen, A. K. (2006) 'What Do We Want from a Theory of Justice?', *The Journal of Philosophy*, 103 (5), 215–38.

Sen, A. K. (2009) *The Idea of Justice*. Cambridge, MA: Belknap Press.

Simmons, A. J. (2010) 'Ideal and Non-ideal Theory', *Philosophy & Public Affairs*, 38 (1), 5–36.

Stemplowska, Z. (2008) 'What's Ideal about Ideal Theory?', *Social Theory and Practice*, 34 (3), 319–40.

Stemplowska, Z. and Swift, A. (forthcoming [2012]) 'Ideal and Nonideal Theory', in D. Estlund (ed.), *The Oxford Handbook to Political Philosophy*. Oxford: Oxford University Press.

Swift, A. (2008) 'The Value of Philosophy in Nonideal Circumstances', *Social Theory and Practice*, 34 (3), 363–87.

Valentini, L. (2009) 'On the Apparent Paradox of Ideal Theory', *The Journal of Political Philosophy*, 17 (3), 332–55.

Wolff, J. (2007) 'Harm and Hypocrisy: Have We Got it Wrong on Drugs?', *Public Policy Research*, 14 (2), 126–35.

Ypi, L. (2011) *Global Justice and Avant-Garde Political Agency*. Oxford: Oxford University Press.