Bridging Ideal and Non-ideal Theory. An Incrementalist Approach to Normative Theorising

Alexandru Volacu

Abstract: In this paper I propose and defend a novel methodological approach for pursuing normative inquiries, which aims to bridge ideal and non-ideal theory and provide a framework for the systematization of knowledge in political philosophy. The core ideas of the approach are that: (1) political philosophy should be concerned with both short term goals, such as addressing severe injustices in the present world and long term goals, such as moving towards gradually more just states of the world in the long-run and (2) an appropriate way to satisfy these aims in a systematic manner is to build a landscape of normative models through incremental derivation (i.e. making the assumptions of a model more or less fact-sensitive one at a time) and selecting to pursue the best path that we might take from mitigating problems of severe injustice in the current state of the world to the achievement of perfect justice. The methodological view proposed features a number of advantages, concerning both the mitigation of problems raised by building normative models exclusively at the level of ideal theory or exclusively at the level of non-ideal theory as well as the transition between ideal and non-ideal theory.

Keywords: desirability, fact-sensitivity, feasibility, ideal theory, normative incrementalism, non-ideal theory, normative model

1 Comments welcomed at: alexandru.volacu.14@politice.ro. Institutional affiliation: National University of Political Science and Public Administration (SNSPA Bucharest), Faculty of Political Science, Department of Political Science and European Studies.
1. Introduction

Discussions on the methodology of political philosophy have in recent years been primarily focused on the ideal/non-ideal theory debate, which encompasses a variety of questions related to the degree to which normative theories are built on the basis of actual empirical facts rather than counterfactuals, the feasibility of implementing normative principles derived from such theories and, more generally, the overall goal of normative theorising. While the debate may extend to the full range of values considered desirable in political philosophy, it has, for the most part, specifically addressed the value of justice, a practice to which I will also adhere in this paper. The positions adopted by various philosophers in this debate are much too numerous to be discussed in this introductory section, but they can broadly be grouped into two camps. On the one hand, authors such as Mills (2005), Farrelly (2007), Sen (2009) or Anderson (2010b) are deeply critical of ideal theory and claim that normative theorising should exclusively take place at the level of non-ideal theory, with a focus on the mitigation of severe present injustices. On the other hand, defenders of ideal theory, such as Robeyns (2008), Stemplowska (2008), Swift (2008), Valentini (2009) or Simmons (2010), do not take an opposite position, by claiming that normative theories should be exclusively built at the level of ideal theory, but instead support the weaker claim that both ideal and non-ideal theory are important for pursuing inquiries in political philosophy. In this paper I will support the latter position, offering a methodological approach for bridging ideal and non-ideal theory, which emphasizes the importance of both levels of theorising, neutralizes some of their shortcomings, offers key insights into the transition from ideal to non-ideal theory and into the transition from non-ideal to ideal theory and provides a framework for systematizing knowledge in political philosophy.

The article is structured as follows: in section 2 I discuss the structure of normative models built at what Hamlin and Stemplowska (2012) have termed the level of institutional design and argue that the evaluative space for such models is represented by their outputs, with inputs playing an important, but only instrumental, role in establishing their overall strength. In section 3 I describe the main issues in contention in the ideal/non-ideal theory debate and

---

2 A general sketch of the main dimensions on which the ideal/non-ideal theory debate has been carried is offered in section 3.

3 I borrow the term from Lawford-Smith (2010, p.367).
argue that they are reducible to a two-dimensional framework, which takes into account idealisations in inputs and idealisations in outputs. I then discuss some of the drawbacks of normative models with a high level of idealisation in inputs/outputs and of normative models with a low level of idealisation in inputs/outputs. In section 4 I provide the outline of a new methodological approach to political philosophy, which I term *normative incrementalism*, by presenting its core ideas and its general structure. I then go on to defend this view by explaining the benefits of pursuing normative inquiries consistent with this framework. Section 5 concludes.

2. Institutional design and the structure of normative models

In the ensuing attempt to examine both the structure of normative theorising and ideal/non-ideal approaches to political philosophy, I will rely on a number of distinctions drawn by Hamlin and Stemplowska (2012) and Stemplowska (2008). The first of these distinctions, proposed by Hamlin and Stemplowska (2012), delineates two different ways of engaging in normative theorising, namely by working at the level of the *theory of ideals*, which seeks to identify, specify the content of values and study their interplay, or at the level of *institutional design*, which seeks to identify and specify the content of social arrangements (in a wide sense, which also includes guiding individual behaviour, not just institutions). While the theory of ideals and institutional design interact to some degree, both are independent fields of normative inquiry. For instance, inquiring whether the best account of welfare is enjoyment, preference satisfaction or an “objective list” (see Arneson: 2000a, pp.511-514) does not presuppose a commitment to any particular institutional design theory, nor does it necessarily lead to a particular proposal at the level of institutional design. In turn, constructing a theory at the level of institutional design, such as Roemer’s (1993) luck egalitarian proposal, is not predicated on

---

4Briefly stated, Roemer’s operationalization of luck egalitarianism entails that "equality of opportunity for X holds when the values of X for all those who exercised a comparable degree of responsibility are equal, regardless of their circumstance" (Roemer: 1993, p.149). This principle is applied in several steps. First, the circumstances which play a central part in shaping the preferences of individuals are identified. Second, individuals are divided into types, with each type consisting of a particular combination of values attached to the circumstances identified in the first phase. Third, a median value is established for each of the types constructed in the second phase. Fourth, differences in the exercise of responsibility are drawn out, by comparing the behaviour of individuals with the median values set in the third phase. Notice that by referring to the median values of a type, degrees of responsibility are comparable even between individuals across different types. To make this idea clearer, Roemer appeals to an example concerning the allocation of healthcare resources for lung cancer contracted from smoking. He identifies occupation, ethnicity, gender, the smoking habits of the parents and income as the relevant circumstances and argues that in a case where one patient “is a sixty-year-old, white, female college professor whose parents stopped smoking when she was seven” (Roemer: 1993, p.151) who smoked for 8 years (which is the median value for her type) and “a second lung cancer victim, also aged sixty, is a black, male steelworker
a specific interpretation of a number of values (for instance in this case, the best account of the currency of distributive justice), although any such theory does seek to instantiate some particular values. In this paper I am primarily concerned with the normative theorising which takes place at the level of institutional design, therefore I will set aside discussions on the structure of inquiries into the theory of ideals.

Call any fully-developed proposal built at the level of institutional design a normative model. A normative model seeks to bring about a specific state of the world by prescribing normative principles which guide the action of individual or collective agents in order to reach that state of the world. In order to describe the structure of a normative model I will appeal to Stemplowska’s (2008) distinction between inputs (assumptions), rules of derivation and outputs (normative principles). Normative model-building therefore consists of three steps: (1) selecting assumptions, (2) constructing rules of derivation and/or deciding on what values the model seeks to instantiate and (3) formulating normative principles. Consider the theory of justice as fairness as an illustrative example. Among the set of inputs (assumptions) employed whose parents were both chain-smokers (Roemer: 1993, p.151), who smoked for 25 years (which is also the median value for his type), we should allocate the same amount of healthcare resources to both patients since they have exercised a comparable degree of responsibility.

I use the term model instead of theory in this case since it provides a better fit with some of the interpretations of the model/theory distinction in the philosophy of science. For instance, in the distinction drawn by Maki (2001, p.9934), “a theory is an abstract and general statement about a number of generic entities and their key dependencies [...]. A theoretical model is a more specific and smaller scale version of the theory, involving various idealisations, simplifications and other auxiliary assumptions, and addressing specific issues. A theory serves as a prototype for the construction of such models”. To illustrate, if we use this terminology, we might say that prioritarianism as a whole represents a theory of justice, while particular operationalizations of the priority view, such as those which can be found in Arneson (2000b), Persson (2001), Brown (2005), Peterson and Hansson (2005) or Dorsey (2014) are normative models. While the normative theory relies on some core thesis (or theses), which for the case of prioritarianism might be that “we ought to give priority to benefitting a worse off group over a better off group whenever (i) the number of people in the worse off group is at least as great as that in better off group, and (ii) the benefit to each of the former is at least as great as the benefit to each of the latter” (Brown: 2005, p.204), the normative model must first satisfy the core thesis, and second, it must give particular interpretations to some of its elements (although not necessarily all; for instance in the case of prioritarianism, specifying the form of the priority-weighted utility function is not always required) in order to make it operational for use in actual or hypothetical cases.

By normative principle I here mean "a general directive that tells agents what (they ought, or ought not) to do" (Cohen: 2003, p.211), where agents can be individuals, institutions etc. By this statement I do not mean to submit to a strictly consequentialist outlook. Instead, the idea of reaching a particular state of the world should be interpreted in a broad sense, as compatible with non-consequentialist frameworks. For instance, a deontological principle which holds that we have a moral duty not to lie may be said to aim at bringing about a state of the world where lying does not occur. I thank Tom Parr for pointing me to this issue.

It is possible to specify both of them or only one of them. For instance, if the rules of derivation are simply those of logical entailment they do not need to be explicitly specified. On the other hand, if the model is of a purely procedural nature, no values are considered as desirable a priori, therefore we would only need to specify the rules of derivation.

Although Hamlin and Stemplowska (2012) hold the view that (at least) Rawls’s first principle of justice belongs to the theory of ideals, assume, for the sake of the argument, that the primary purpose of justice as fairness is to guide institutional design.
we find: a) moderate scarcity, b) strict compliance, c) favourable conditions, d) maximizing rationality, e) mutual disinterest and so forth. The rule of derivation consists in parties unanimously choosing principles of justice behind the veil of ignorance\textsuperscript{10}. The outputs generated are the two principles of justice as fairness, arranged in serial order.

The strength of a normative model can be assessed in several ways. One of the most common methods employed in contemporary political philosophy is to examine specific aspects of the state of the world to which normative principles give rise to and contrasting these implications with one’s own moral intuitions. Following Sen (1979, p.197) we may call this method the \textit{case-implication approach}. A second way of assessing a normative model is by examining the consistency between the principles generated and other principles or values which may be thought to be more fundamental. Again, following Sen (1979, p.198), we may call this the \textit{prior-principle approach}. Since both of these approaches involve substantive moral judgements about the outputs of a model, we might collectively call them \textit{desirability} evaluations\textsuperscript{11}. A further way to assess a normative model is by examining the feasibility of normative principles. Call this a \textit{feasibility} evaluation. There are at least two distinct strands of feasibility considerations related to outputs. On the one hand, we might say that an output is more or less feasible to the extent that the normative principles generated are more or less difficult to implement. For instance, a model instantiating equality of welfare through institutional redistribution would be relatively unfeasible in the present state of the world since it would require us to obtain information about individual resource-to-welfare conversion rates which seem to be directly unavailable to us. On the other hand, we might say that an output is more or less feasible to the extent that the move from the status-quo to the possible state of the world is more or less difficult to achieve. For instance, a model which seeks to achieve a state of the world with no poverty through individual contributions for poverty eradication would be relatively unfeasible in the present state of the world since many individuals would presumably not provide their fair share of contributions. Furthermore, an important distinction on the topic of feasibility is advanced by Gilabert and Lawford-Smith (2010), who distinguish between \textit{hard} constraints (e.g. logical, nomological, biological or many types of psychological constraints), that are not liable to being changed (at least in the foreseeable future), from \textit{soft} constraints (e.g. economic, institutional or cultural), which "place limits on what people are

\textsuperscript{10}Justice as fairness is a special case of normative model, since the rules of derivation actually involve the use of a fairly complex mechanism. As specified in a previous footnote however, most normative models simply rely on logical entailment as a rule of derivation.

\textsuperscript{11}McDermott (2008, p.17) considers that these type of assessments are analogous (though not similar) to empirical tests in scientific fields of inquiry.
comparatively more likely to do, but the limits are neither permanent nor absolute” (Gilabert and Lawford-Smith: 2010, p.813). Both the distinction between desirability and feasibility and their joint importance are emphasized, among others, by Raikka (1998), Gilabert (2008), Stemplowska (2008), Hamlin and Stemplowska (2012) and Carey (2015).

The previously described methods of evaluating a normative model concerned its outputs. Thus, a second method of evaluation might concern the inputs of the model. Let us consider that an assumption is: (1) metaphysically impossible, if there is no possible world in which it can be true, (2) physically impossible, if it cannot be true for the present world but is not impossible in every possible world, (3) empirically false, if it can be true for the present world but is currently false and (4) empirically true, if it is true for the present world. The assumptions with which we are usually concerned in constructing and assessing normative models fall within the latter two categories. Call a model which has as an input an assumption X which is empirically true, fact-sensitive with respect to X and one which has as an input an assumption X which is empirically false, fact-insensitive with respect to X. As Farrelly (2007, p.847) and Hamlin and Stemplowska (2012, p.51) suggest, normative models can be represented on a fact-sensitivity continuum, with complete fact-insensitivity at one end and complete fact-sensitivity on the other end. While Farrelly places Cohen (2003) and Mason (2004) on the extreme end of fact-insensitivity and no model on the extreme end of fact-sensitivity, a more appropriate claim would be that all models belong to intermediate positions on this continuum, since no model would normally rely on metaphysically impossible assumptions, almost no model relies on physically impossible assumptions and it is difficult to conceive of a model where no assumptions at all (even if only implicit) are empirically true. Also, in models that seek to be action-guiding, some of the assumptions about the state of the world...
world must be true, since the model’s aim is to change a particular aspect of the state of the world (for instance, a model which seeks to mitigate racial discrimination must start from a state of the world where it is true that racial discrimination exists as part of its assumptions set). Furthermore, no model is fact-sensitive for every assumption, since at the very least it needs to make some abstractions or idealisations\textsuperscript{17} in order to be operational.

Some political theorists argue that fact-sensitivity can be directly employed as a criterion in distinguishing good normative models from bad ones. One of the advocates of this position is, in fact, Farrelly (2007, pp.846-848), who claims that political theorists that occupy a moderately ideal position, by which he means that they only adopt a fairly narrow range of empirically-true assumptions, are too fact-insensitive and should move towards non-ideal theory, where a wide range of assumptions are empirically true (although not all, for the purposes of avoiding the simple reiteration of the status-quo). I hold that this view is mistaken. While it might sometimes be the case that a model which is more fact-sensitive is better than one which is less fact-sensitive, this conclusion cannot be drawn simply on the basis of looking at the sensitivity to empirical facts of inputs, since they only play an\textit{instrumental} role in the model. Rather, the evaluative space for a normative model exclusively concerns its outputs\textsuperscript{18}, with models being hierarchically ranked according to the combination of desirability and feasibility yielded\textsuperscript{19}. Farrelly’s own objections to Rawls (1971) and Dworkin (1981) do not challenge this instrumentalist view, since his claim is that when we are making the Rawlsian and Dworkinian models incrementally more fact-sensitive, the outputs of these models change. But, assuming that he has correctly judged the impact which his assumption modifications have on the outputs, then the evaluative space used is still connected exclusively to outputs. At least in part, the disagreement may arise from a conflation of the idea of fact-sensitivity and feasibility, which seems to be present in Farrelly (2007). The two issues are distinct and

\textsuperscript{17}See section 3 for the distinction between the two.

\textsuperscript{18}This substantive position, which I do not mean to further defend here, is analogous to the empirical-instrumentalist view in the philosophy of economics (see Friedman: 1966; MacDonald: 2003) which states that the evaluation of models should only take into account their predictive success, not the truthfulness of their assumptions. Since the discussion targets normative, not positive models, predictive success is of course not the object of concern here, but the exclusive focus on the outputs of models still stands.

\textsuperscript{19}I do not intend to address the question of how desirability and feasibility should be measured or traded-off (see Carey: 2015, pp.153-154 for some preliminary suggestions). In fact, it is possible that no consistent way of engaging in this trade-off exists. However, even if it turns out that the two values are incommensurable, this would not lead to a complete indeterminacy in the evaluation of normative models. First, if a model outperforms another one in terms of\textit{both} desirability and feasibility, we may claim (borrowing from the language of game theory) that the former\textit{dominates} the latter and is always preferable, without assuming anything about the way in which desirability and feasibility are traded-off. Second, we can also avoid the trade-off by assigning absolute priority to one of the two values, claiming that the other may only come into play (if at all), when two normative models perform equally well on the dimension which takes priority.
applicable to different components of a normative model, with fact-sensitivity being a property of inputs and feasibility a property of outputs. It might sometimes be the case that a higher degree of fact-sensitivity indeed leads to a more feasible output. For instance, take a normative model that seeks to instantiate a sufficientarian policy through institutions that constrain individuals to give a fair share of their income for poverty eradication. Also, assume non-strict compliance and that individuals are egoists. Weakening the assumption of egoism and conceptually allowing for individuals to also be influenced by altruistic considerations (an assumption which is more fact-sensitive than that of strict egoism) might make the proposed outputs more feasible, in that there is a higher likelihood that the prescribed state of the world would be achieved. However, it might also be the case that increasing fact-sensitivity leads to a decrease in the feasibility of the outputs. Take again Roemer’s (1993) luck egalitarian normative model and his lung cancer example. Assuming that only occupation matters in the process of developing a smoking behaviour would yield a fairly straightforward implementation of the proposed policy. However, as we incorporate more fact-sensitive assumptions (parental attitudes towards smoking, environmental conditions during childhood etc.) which are to some extent relevant for the development of smoking behaviour, the policy becomes less feasible, since the normative principles resulted are harder to implement. Finally, there may be other assumptions which play no part in influencing the feasibility of a model. Thus, fact-sensitivity and feasibility come apart, and while they do interact in some cases, the type of interaction is neither necessary, nor does it always point in the same direction.

However, even if fact-sensitivity is not directly usable for evaluating normative models, there are a number of ways in which fact-sensitivity might prove to be of considerable, albeit indirect, importance, through its interaction with the outputs of the model. At least four such possibilities come to mind Firstly, it might be the case that modifying an assumption by incrementally making it more fact-sensitive than in the original model might generate different normative principles than the ones which had initially resulted from the model. Secondly, it might be the case that under certain assumptions, the state of the world which the model tries to bring about through its outputs is highly desirable, but when we modify the status of one of the assumptions by making the model more fact-sensitive, we notice that the state of the world brought about by the model is much less desirable. Thirdly, it might be the case that modifying an assumption by incrementally making it more fact-sensitive than in the original model might make the normative principles generated less feasible. Fourthly, it might also be possible that critiques raised against a model might draw their force from fact-insensitive assumptions and
incrementally moving in the direction of fact-sensitivity might undermine such objections altogether. Thus, understanding the exact way in which fact-sensitivity, as a property of the set of inputs, affects the outputs of a normative model is of paramount importance for the ideal/non-ideal theory debate and, as further explained in section 4, represents a strong reason in favour of the methodological approach advanced in this article. In the subsequent section I set the background for the approach by reformulating the ideal/non-ideal theory distinction and presenting some of the drawbacks of each type of view.

3. On the ideal/non-ideal theory distinction. A two-dimensional reconstruction

As mentioned in the introductory section, recent discussions on the methodology of political philosophy have largely focused on whether normative models should be constructed at the level of ideal or non-ideal theory. The distinction between ideal and non-ideal theory, like many other on-going debates in political philosophy, can be traced back to Rawls (1971). In its initial formulation, ideal theory starts from an assumption of "strict compliance and works out the principles that characterize a well-ordered society under favorable circumstances" (Rawls: 1971, p.216), while non-ideal theory takes into account both partial compliance and unfavourable circumstances, opening up discussions pertaining to civil disobedience, just wars etc. Even though the Rawlsian cut has often been seen as being fairly clear and straightforward, it has recently been critically appraised and expanded upon by Estlund (2008), Simmons (2010), Arvan (2014) and Chabhoun (2015). However, in the current landscape of political philosophy the meaning of ideal and non-ideal theory does not necessarily refer to the original one, as introduced by Rawls.

While a recently salient definition has been advanced by Stemplowska (2008, p.324), who holds that non-ideal theories issue recommendations that are both desirable and achievable, while ideal theories lack this feature, the contrast between ideal and non-ideal theories, as these terms have been used in the past years, cannot be ultimately captured in any one-dimensional framework. Thus, Hamlin and Stemplowska (2012, pp.48-52), identify four dimensions (interpreted as continuums) on which this distinction has been construed. The first of these dimensions is the classical Rawlsian one, which distinguishes between full and partial

---

21 As Simmons (2010, pp.5-6) points out.
22 See Jubb (2012) for a call to a return to the original, Rawlsian, meaning of the distinction between ideal and non-ideal theory.
23 This definition is employed, for instance, by Carey (2015).
compliance, and has been briefly discussed above. This distinction has been used, for instance, by Eddy (2008) in her analysis of ideal rights. The second dimension distinguishes between idealisations and abstractions. This cut has been proposed by O'Neill (1987) and concerns the difference between bracketing some features of the state of the world in order to engage in normative theorising (abstractions) and assuming falsehoods about these features (idealisations). This distinction has been used, for instance, by Mills (2005) in his critique of the ideological charge of ideal theory and by Erman and Moller (2013) in their defence of ideal theory against various objections levied against it. A third dimension distinguishes between fact-sensitivity and fact-insensitivity. This distinction is usually attributed to Cohen (2003) and while Cohen himself uses it to delineate fundamental normative principles (which are fact-insensitive) from principles of social regulation (which are fact-sensitive), in the context of the ideal/non-ideal debate it usually refers to the degree to which theories incorporate factually accurate or counterfactual assumptions. This distinction has been used, for instance, by Farrelly (2007) in his critique of Rawlsian and Dworkinian theories of justice. Finally, the fourth dimension distinguishes between transcendental and comparative theories. The distinction is traceable to Sen (2009) and concerns the goals of normative theories, which can be either to focus on the achievement of the perfect instantiation of justice or to focus on remediating the most egregious instances of injustice. This distinction has been used, for instance, by Schmidt (2011), Freeman (2012), and Robeyns (2012) in their analysis of ideal and non-ideal theory.

How does the ideal-non/ideal theory debate, outlined above, map unto the framework presented in the previous section? The claim which I mean to advance in the ensuing discussion is that the various dimensions identified can be rigorously reduced to just two, namely (1) input idealisation and (2) output idealisation. Specifically, the first three dimensions outlined above belong to the category of input idealisation, while the fourth one belongs to the category of output idealisation. To see that this is the case, let us examine each of the distinctions in turn. First, the Rawlsian delineation of ideal from non-ideal theory centrally relies on the presence

---

24 Although, as Stemplowska and Swift (2012) rightly point out, Cohen's arguments are not about the merits of ideal theory, being, at often times, mistakenly entangled in these debates.

25 Or, more sharply stated by Hamlin and Stemplowska (2012, p.51), between theories of perfect justice (or other values) and theories of local improvement in justice (or other values)

26 Slightly different ways of spelling out the distinction between ideal and non-ideal theory have also been proposed by Valentini (2012), who distinguishes between the full compliance/partial compliance dimension, a utopian/realistic dimension (which largely mirrors the fact-sensitivity dimension) and a transitional/end-state dimension (which largely mirrors the transcendental/comparative dimension) and by Phillips (2008), who distinguishes between the Rawlsian cut, the idealisation/abstraction dimension and the fact-sensitivity dimension. I take the distinction proposed by Hamlin and Stemplowska (2012) to be more useful for the purpose of drawing out the contrast between ideal and non-ideal theory, since it encompasses a wider region of the debate than Valentini and Phillips's accounts, taken in isolation.
of a specific assumption, namely that individuals are fully compliant with the principles of justice\textsuperscript{27}. Whatever special place this assumption may have for the ideal/non-ideal theory continuum, it is formally still only an assumption about the state of the world, akin to other assumptions which together constitute the set of inputs of a normative model (for instance the assumption that individuals are utility maximizers). Second, the idealisation/abstraction dimension can be drawn in two different ways, starting from O'Neill's original definition. The first interpretation, which I think is most faithful to O’Neill (1987), is that idealisations involve making false assumptions about a fact while abstractions involve not assuming anything about a fact. The second, and I think more interesting way of drawing it, is suggested by Hamlin and Stemplowska (2012, p.50), and refers to the significance which the assumption has on the conclusions of the theory. While Hamlin and Stemplowska do not themselves suggest that the two interpretations may pull in different directions, we can use various examples to show that these is the case. A false assumption can conceivably play no part in the outputs of a model. For instance, Rawlsian justice as fairness is fully robust to a change from the neoclassical account of rationality employed by Rawls to Levi’s (1984) conception, as shown by Angner (2004). In this case, it would be an idealisation in the first sense but an abstraction in these second sense, if we accept the premise that Levi’s conception of rationality is more empirically plausible than the neoclassical account. Conversely, a fact about which the model makes no assumptions might be significant for the outputs, since taking it into account might prove to change the normative principles generated or increase/decrease its desirability or feasibility. It would therefore be an abstraction in the first sense, but an idealisation in the second sense. If we follow the interpretation suggested by Hamlin and Stemplowska then, the idealisation/abstraction (or lack of idealisation) distinction can also be captured in the framework of fact-sensitivity, with idealised assumptions being \textit{significantly fact-insensitive} and abstractions being \textit{insignificantly fact-insensitive}, while non-idealised assumptions are simply fact-sensitive ones. Once again then, under this interpretation of the idealisation/abstraction dimension, the object of concern is the idealisation of the set of inputs of the normative model. Third, the distinction between fact-sensitivity and fact-insensitivity falls in this category in a direct way, since fact-sensitivity is a property of inputs. Finally, the transcendent/comparative distinction can be understood to refer to the outputs of a normative model, with models that place a greater weight on desirability than on feasibility being located

\textsuperscript{27}See for instance Simmons's (2010, p.8) statement that "the defense of the ideal theory of justice is [...] to proceed against one straightforwardly unrealistic background assumption: that of 'strict compliance'".
closer to the ideal theory extreme and models that place a greater weight on feasibility being located closer to the non-ideal extreme.

Thus, as I illustrate in Figure 1, the ideal/non-ideal distinction can be mapped on to a two-dimensional continuum, with input idealisation as one dimension, in which the extreme points are represented by fact-sensitivity or insignificant fact-insensitivity for non-ideal theory and significant fact-insensitivity for ideal theory and output idealisation as another dimension, in which the extreme points are represented by the absolute priority\(^{28}\) of feasibility over desirability for non-ideal theory and the absolute priority of desirability over feasibility for ideal theory\(^{29}\).

Fig. 1. Two-dimensional characterization of the ideal/non-ideal distinction

I will now attempt to explain why normative models built at the level of both ideal theory and non-ideal theory\(^{30}\) are important on their own, but suffer from various shortcomings when

---

\(^{28}\)By the absolute priority of desirability/feasibility I mean that the goal of the normative model takes exclusive account of that specific feature. Normative models that fall between the two extremes are understood to give relative priority to one of the two features, in the sense that they place differentially greater weight on either desirability or feasibility.

\(^{29}\)To some extent, the two-dimensional characterization of the ideal/non-ideal distinction proposed here mirrors the two senses which Valentini (2009, pp.337-338) attributes to ideal theory, namely the non-technical sense, which refers to theories that aim to advance ideals of fully just societies, and the technical sense, which refers to building theories under idealised assumptions. However, the cut which I propose here is more comprehensive since it also takes into account the significance of fact-insensitivity, not just its presence, and it also spells out the relevant features for the assessment of the degree of idealisation in normative principles.

\(^{30}\)I use these labels for purposes of terminological simplification. However, the labels “more-ideal theory” and “less-ideal theory” used by Hamlin and Stemplowska (2012) are clearly more accurate in this framework.
looked at in isolation from one another. Let us first look at ideal theory, and more specifically at normative models which maintain a high degree of idealisation in outputs. The main advantage of models of this type is that they are able to offer normative principles that can bring about highly desirable states of the world and with which we seem to be ultimately concerned when engaging in normative theorising. The main drawback of such models, however, comes in the form of a standard critique raised against ideal theory, namely that it fails to be action-guiding for the real world (see Valentini: 2009, pp.340-343 for an overview of three possible strands of this critique). This objection loses some of its force under the view assumed here however, which follows Hamlin and Stemplowska (2012) in distinguishing between the theory of ideals and institutional design, since in this interpretation normative models (even ideal ones) built at the level of institutional design are by definition action guiding. Still, to the extent that they are relatively or even absolutely unconcerned with questions of feasibility, it may be possible that the models are action-guiding only under very particular states of the world, and will prove unable to tell us how to mitigate injustices in the "real" world31, or otherwise stated, about moves that we might take from the current state of the world to a less unjust, but feasible, state of the world. A different problem afflicts normative models that maintain high levels of input idealisation. The usefulness of such models relies on the fact that they are able to offer guiding principles in cases where the state of the world that is taken as the status-quo is characterized by a set of counterfactual assumptions which can be presumed to be more conducive in instantiating the value considered important in the model, e.g. the assumption of strict compliance in normative models concerning justice. The main drawback, however, is that such models may not have much to say about moves from current states of the world either, since the alterations of the set of inputs may lead to alterations in the set of outputs in (at the very least) all the ways described at the end of section 2. Thus, knowing that under a highly idealised set of assumptions a normative model will generate a particular account of principles of justice may turn out to be useless for realising actual improvements in justice, since the normative principles resulting under high level of input idealisation might not be the same as the ones resulting under lower levels of input idealisation, might be unfeasible in the current state of the world or may turn out to be much less desirable for the current state of the world.

31 Note that I do not mean to make the claim that ideal theory cannot tell us anything relevant about mitigating injustices in the current state of the world, since it may still be useful in shaping the moral convictions through which we assess present injustices. However, I believe it is plausible to hold the weaker claim that if we only engage in ideal theory we might miss out on some important features which contribute to the injustice of the current state of the world, due to the fact that we significantly idealise various assumptions about the world.
Non-ideal theory escapes the problems previously underlined, but pays a non-trivial price in order to do so. First of all, theories that maintain a low degree of idealisation in outputs are vulnerable to the important objection raised by Simmons, who claims that "we, as theorists of justice, simply should not care which policies are politically possible (etcetera) unless those policies are also on an acceptable path to a just institutional structure" (Simmons: 2010, p.34). Thus, without having a conception of the most desirable accounts of justice, which are developed at the level of ideal theory, non-ideal normative models might mitigate some particular injustices occurring in the current state of the world but lead us away from achieving a more just state of the world in the long-run. Further, the more normative models move to the extreme where feasibility takes absolute priority over desirability, the more vulnerable they become to the possibility of failing to seek the remediation of more severe injustices, if the remediation of less severe injustices is somewhat more feasible. Normative models that maintain a low degree of idealisation in inputs are also met with a different type of problem. Since such models avoid (as far as possible) to include significantly fact-insensitive assumptions in their set of inputs, they are unable to detect what soft constraints (see Gilabert and Lawford-Smith: 2012, p.813) might make a more desirable state of the world less feasible and might be themselves removed if acted upon. Thus, non-ideal normative models are bound to treat some features of the current state of the world as fixed, even though this might not be the case, and therefore fail to recognise opportunities to move to more just outcomes, if these features simply embody soft constraints.

4. Normative incrementalism as an approach to political philosophy

In this section I propose a particular methodological approach, which I term normative incrementalism, that seeks to offer an image of the way in which we should engage in normative theorising. The section proceeds in three steps. I will first present the view in its most general form, disregarding questions related to complexity which will inevitably arise in its implementation. Second, I will present the advantages of pursuing the methodological view outlined here, which pertain both to the mitigation of the problems raised by exclusively theorising at the ideal or the non-ideal level (discussed in the previous section), but also to the transition from non-ideal to ideal theory and the systematization of knowledge in political philosophy. Third, I will argue that these upshots are not eliminated even if we are only able to use normative incrementalism in order to provide a partial picture of the fully conceivable landscape of normative models.
Normative incrementalism starts from the idea that political philosophy should give adequate weight to both the elimination of pressing injustices in our current state of the world and at the same time to moves towards considerably more just states of the world in the medium and long-run. It then advances the claim that in order to achieve this we need to construct a comprehensive image of the various intermediary states of the world, and the normative principles which bring them about, placed between the current state of the world and one which embodies perfect justice. Finally, it maintains that this image is most appropriately constructed by starting from a particular normative model and move to both ideal and non-ideal theory in gradual steps\textsuperscript{32}, by making the set of inputs incrementally more or less fact-sensitive. Call the normative model from which we incrementally derive subsequent models a parent-model, the models which are derived from the initial one offspring-models and the picture obtained by aggregating parent-models and offspring-models a landscape of models. The sequence which I have in mind for constructing a model-tree is therefore something of the following sort\textsuperscript{33}: (1) the values sought to be instantiated and/or the rules of derivation in use are fixed\textsuperscript{34}; (2) an account (as extensive as possible) of the set of assumptions (fact-sensitive or fact-insensitive) considered significant for the purposes of the normative model is offered; (3) an original parent-model, which starts from a particular combination of idealised and non-idealised assumptions in the set of inputs, is constructed; (4) the original parent-model is incrementally altered by changing assumptions one at a time, making them more or less fact-sensitive. Successively, offspring-models become parent-models for newer, incrementally-altered models; (5) with a relatively well-developed landscape of models, comparative assessments of models in terms of desirability/feasibility are performed, as well as assessments of the paths from current feasible alternatives to the most desirable ones through incremental changes in facts about states of the world, seeking to select the path which maximizes an overall desirability/feasibility combination.

\textsuperscript{32}I do not take any particular view on the question of the level where original models should be constructed. Thus, the approach is compatible both with a position which would state that we should first build models at the level of non-ideal theory, in order to expedite the amelioration of present injustices and with a position which would state that we should first answer questions about perfect justice and then move to uncover normative principles which are feasible for the current state of the world.

\textsuperscript{33}Importantly, this is only one way of operationalizing the core idea of normative incrementalism. Rejecting the particular sequence proposed does not necessarily lead to a rejection of the approach altogether.

\textsuperscript{34}I assume that discussions of what value or set of values should be instantiated and the adequacy of rules of derivation takes place at the level of the theory of ideals. Thus, disagreement on the question of which values should the idea of justice embody carries over to the field of institutional design, in that each configuration of values should yield a particular normative landscape. In one way, the respective values can also be understood as inputs of a normative model, but they constitute a normative type of assumptions, which are fixed for the purpose of performing incremental derivations. Different rules of derivation can also be understood, in the same manner, as giving rise to different landscapes of normative models.
This latter claim states that we should not simply pursue the shortest path to perfect justice (even if that would lead to severely unjust intermediate states of the world), but that the path selected must itself be the one that best instantiates justice overall. The claim is substantive and independent of the normative incrementalist framework as such and also departs from Simmons’s (2010) original argument in favour of the role of ideal theory, which did not focus on the path as an important object of assessment in itself, but only on its instrumental relation to perfect justice. However, the case that we should not simply look at the path from the current state of the world to the most just state of the world with a focus on the shortest route can be supported by both deontological and consequentialist reasons. On deontological grounds, it would be wrong to discard the injustice occurring in intermediate states of the world, since agents populating those states of the world are of identical moral relevance as agents in the perfectly just state of the world. Thus, if we would be able to achieve perfect justice in 100 years by severely curtailing liberty and equality for the next generation, deontological considerations would prescribe rejecting the pursuit of this route since we would violate moral constraints that prohibit us from treating individuals as means for the benefit of others. On consequentialist grounds, it would be wrong to discard the injustice occurring in intermediate states of the world, since the disvalue brought about in these intermediate states might be sufficient to outweigh the value brought about by perfect justice. Building on the same example, consequentialist considerations might prescribe rejecting the route if the disvalue brought about by injustices which affect the next generation would be so great that they would diminish the overall value brought about by perfect justice, in comparison with an alternative path where perfect justice might be achieved within a longer timeframe (or perhaps not at all) but the disvalue of injustices in intermediate states of the world would be much less severe.

It is probably quite clear by now that the general view outlined above may be unworkable if seeking to map out the whole set of possible assumptions about states of the world. Even in highly simplified examples, such as one where we would have only three assumptions we would obtain eight sets of outputs by taking into account all possible

---

35In the sense that normative incrementalism might accommodate other substantive ideas concerning institutional design, e.g. that we should exclusively be concerned with reaching perfect justice in the shortest possible time, regardless of the injustices occurring in intermediate states of the world.

36This argument mirrors, to some extent, Buchanan’s proposal of placing a constraint of moral accessibility on ideal theorising, through which he means that “the transition from where we are to the ideal state of affairs should be achievable without unacceptable moral costs” (Buchanan: 2004, p.61).

37Although, granted, some of them might be identical.
combinations of assumptions\textsuperscript{38}, and in real models we have many more assumptions to take into consideration which lead to an exponential increase in the sets of outputs. Furthermore, disagreements at the level of theory of ideals concerning the content of justice\textsuperscript{39} and disagreements concerning the most adequate rules of derivation would lead to monumentally and, most likely, irreducibly complex landscapes of normative models in political philosophy. An objection against normative incrementalism might then state that a complete picture of these landscapes are therefore bound to be unattainable and that the view itself fails as a consequence. While the first part of the objection is surely correct, since mapping out all possible normative models would require identification of all the features which characterize a state of the world at some point and all the ways in which these features might feasibly be changed, this does not entail the wholesale rejection of normative incrementalism. To justify this latter position, I will underline the positive upshots of adopting this methodological approach and show that these upshots also stand if we are only able to construct a partial (but relatively well-developed) landscape of normative models\textsuperscript{40}.

First, consider the main problems discussed in relation to building normative models exclusively at the level of ideal theory, mentioned in the previous section. On the one hand these models can fail to be action-guiding for the current state of the world due to a high level of output idealisation. On the other one, the normative principles generated under significantly fact-insensitive assumptions, due to a high level of input idealisation, might be different, less feasible or less desirable than the ones of a similar model but which is constructed based on a wider range of empirically true assumptions\textsuperscript{41}. Normative incrementalism mitigates the former

\textsuperscript{38}And even more than that if we would take into account the full range of ways in which assumptions can be empirically false.

\textsuperscript{39}For instance, if we take a value-pluralist approach which considers that the interplay of values takes place within a theory of justice, diverging proposals may be advanced at the level of the theory of ideals on the proper weight which we should assign to each value which carry over to disagreements at the level of institutional design as well.

\textsuperscript{40}Alternatively, it could be stated that even if the view would perhaps turn out to be too complex to be fully implementable, it does not mean that we should discard it as incorrect. Borrowing from Swift’s (2008) terminology, it could be said that while it may fail as a \textit{practical} conception of normative theorising, it might still resist as an \textit{epistemological} conception. However, since I wish to suggest that the view itself has a methodologically guiding role and does not therefore fail as a practical conception, I will not rely on this point to make the case in its favour.

\textsuperscript{41}Normative incrementalism is not committed to any substantive view on what constitutes a bad idealisation of inputs. For some proposals on this issue see Robeyns, who argues that "bad idealisations are those that do not serve to model a certain absence of injustice at the level of ideal theory in cases in which this is theoretically justified" (Robeyns: 2008, p.358) and that "bad idealisations amount to ignoring the existence of certain forms of injustice that need to be theorized rather than simply ignored in the theorizing" (Robeyns: 2008, p.358), Valentini, who argues that bad idealisations are those that "build into a normative theory a false, i.e., idealised, account of the social phenomena the theory itself aims to put under moral scrutiny, in this way severely undermining its potential for guiding action in the real world" (Valentini: 2009, p.352) or Uberti, who argues that a badly idealised
problem by requiring that we provide an extensive landscape of models, some of which would provide outputs that are feasible for the current state of the world and the latter problem since it is able to pinpoint what assumptions are significantly fact-insensitive and in what way are normative principles dependent on these type of assumptions.

Second, consider the main problems discussed in relation to building normative models exclusively at the level of non-ideal theory, also mentioned in the previous section. On the one hand, by focusing on feasibility at the expense of desirability, these models may both lead us away from pursuing the best path for instantiating justice in the long-run and at the same time may lead us to endorse normative models that are more feasible but would not eliminate many pressing injustices over somewhat less feasible, but much more desirable, models. On the other one, by only taking into account models developed largely under fact-sensitive or insignificantly fact-insensitive assumptions, non-ideal models are unable to provide a picture of the nature of the constraints that make certain states of the world unfeasible, some of which could technically be removed. Normative incrementalism responds to all these issues, by requiring that we comparatively assess various justice-improving paths and select the one which offers the best overall combination of desirability/feasibility and that we take into account a wide range of both ideal and non-ideal models, some of which will include significantly fact-insensitive assumptions that embody soft constraints and may be changed.

Third, normative incrementalism offers a conceptual framework for the bidirectional transition between ideal and non-ideal theory. The transition from ideal to non-ideal theory (see Phillips: 1985; Robeyns: 2008; Herzog: 2012; Heath: 2013) primarily refers to the way in which we can make use of the normative principles developed at the level of ideal theory in order to address injustices in the current state of the world. As previously mentioned, by mapping out incrementally different models built both at the ideal and the non-ideal level, the methodological approach advanced here allows us to detect the various ways in which principles generated by highly idealised models would be affected by facts constraining the current state of the world. By contrast, the transition from non-ideal theory to ideal theory (see Simmons: 2010; Hendrix: 2013) primarily refers to the idea that the construction and implementation of normative principles designed to mitigate present injustices should be constrained by the pursuit of a path toward achieving perfect justice in the long-run. Once again, normative incrementalism offers an appropriate framework for pursuing this transition,

theory "is one such that it is either practically impossible to rectify the real-world injustices it assumes away, or (more likely) one that rests on false and unrealizable factual premises" (Uberti: 2014, p.223).
since it seeks to map out a landscape of normative models within which we can draw and compare various paths for moving from mitigating injustices in the current state of the world to achieving perfect justice.

Fourth, normative incrementalism can also be interpreted as a tool for systematizing knowledge in the institutional design field of political philosophy. The standard contemporary practice for developing normative models often takes one of the following forms:\(^{42}\): (1) the author begins with a critical appraisal of an existing model or theory and shows that it has some undesirable features (e.g. it has deeply counterintuitive moral implications in some cases or it offers principles that conflict with some fundamental values), subsequently proposing a model that avoids these undesirable features, (2) the author begins with a critical appraisal of an existing model or theory and shows that it focuses on the wrong value or combination of values, subsequently proposing a model which focuses on the "appropriate" value or combination of values, (3) the author identifies a particularly pressing injustice and seeks to provide principles for mitigating the perceived injustice. Undoubtedly, important advances can be made in political philosophy by pursuing either of these approaches. But the relative isolation in which they are normally constructed precludes systematic evaluations of these models, since they are either compared with only one or a handful of other models and on only one or a handful of relevant characteristics (for instance, the desirability of the response which both models offer in a particular counterfactual case)\(^{44}\), or they focus on targeting one particular injustice, potentially missing out on the way in which the normative principles proposed might deepen other important patterns of injustice.\(^{45}\) Thus, by constructing a landscape of normative models we would be able to make some progress (arguably, more than has been realised thus far) in purging models from their relative isolation and offer a comprehensive framework for wide-scope comparison with other models on both desirability and feasibility accounts. Furthermore, in relation to the systematization of knowledge, the incrementalist approach could also provide a valuable way of uniting various theoretical developments that have broadly similar targets but are not construed in a particular relation with one another. For instance, a landscape of models using Rawlsian justice as fairness as an original parent-model could bring together a number of normative models, such as those put forward by Angner (2004), Farrellly (2007) or

---

\(^{42}\)I intend for the list to play only an illustrative role, therefore I do not seek to offer an exhaustive set of possible approaches.

\(^{43}\)This latter approach is especially met at the level of non-ideal theory. Anderson (2010b) can be considered a paradigmatic case for this type of approach.

\(^{44}\)This is the case with the first two methods described above.

\(^{45}\)This it is the case with the third method described above.
Arvan (2014), that build their cases precisely on the sort of incremental derivation which the methodological approach proposed here advocates (without, of course, explicitly adhering to this terminology), thereby enabling comparisons between them and contributing to the systematic accumulation of knowledge within the field of institutional design.

The advantages underlined in this section do not evaporate even if we will not be able to give a complete account of the landscape of normative models that might be constructed starting from a particular rule of derivation or from the instantiation of a particular value or set of values. For one, the approach takes note of the urgency of building models which possess a high degree of feasibility, in order to move from current states of the world to more desirable ones. Even if the list of possible models (and, consequently, moves) will be incomplete, it would still be able to offer principles for guiding action here and now. Additionally, even if we will not be able to map out the whole range of ways in which highly idealised models would fare under the circumstances of the current state of the world, the various intermediate models which are incrementally derived from more ideal ones will still be able to tell us important things about the robustness of these models and about the desirability or feasibility of applying idealised normative principles under present circumstances. Furthermore, even though we will not be able to map out the complete set of paths on which we could theoretically set out from our current state of affairs towards perfect justice, we will still be able to draw at least a partial picture of such paths and perform comparative assessments between them. While, admittedly, this will result in an imperfect evaluation of the various ways in which we could pursue the instantiation of gradually more just states of the world, it would no doubt constitute an improvement over theorising without any consideration whatsoever of the overall consequences (on both the short and long term) that implementing certain normative principles may have, and ultimately, might be the best we could hope for in this regard. Finally, the idea that normative incrementalism would be an appropriate instrument for systematizing knowledge in political philosophy requires no further particular defence, since this role does not imply, from the outset, that the picture resulting from the aggregation of normative models should be a complete one.

5. Conclusions
In this article I set out to offer the preliminary outline of a novel methodological approach to normative theorising at the level of institutional design, which I term normative
incrementalism, and in the course of pursuing this goal I have also addressed some questions related to the structure of political philosophy and the ideal/non-ideal theory debate. Normative incrementalism embodies two core ideas. First of all, it maintains that political philosophy should be concerned with both short term goals, such as mitigating pressing injustices in the present world and long term goals, such as gradually moving towards ever more just states of the world. Secondly, it maintains that an appropriate way to satisfy both these goals in a systematic manner is by constructing a landscape of normative models through incremental derivation of newer models from older ones (i.e. by making a single assumption more or less fact-sensitive) and selecting to pursue one of a number different paths which we might take from tackling problems of severe injustice in the current state of the world to the achievement of perfect justice, according to certain criteria. While in this paper I advocate such a criterion, namely the maximization of overall desirability/feasibility for the states of the world instantiated throughout a specific path, this criterion is not necessarily required by the incrementalist approach. A number of advantages of the approach proposed have been highlighted the previous section: (1) it avoids many of the important problems raised by constructing models exclusively at the level of ideal theory (e.g. that they might fail to be action-guiding for the present world) or non-ideal theory (e.g. that they are unconcerned with possible deviations from the most appropriate paths to the achievement of perfect justice), (2) it provides key insights for operationalizing both the transition from ideal to non-ideal theory and for the transition from non-ideal to ideal theory and (3) it offers a framework for the systematization of knowledge in political philosophy\textsuperscript{46}.

Naturally, since the view I advance in this paper seeks to address a number of different foundational problems of methodology in political philosophy simultaneously and, in the course of this endeavour, to offer a fundamental framework for thinking about political philosophy and pursuing normative theorising, it is most likely vulnerable to a host of objections. Thus, perhaps the best which can be hoped for might just be that the sketch of

\textsuperscript{46} Aside from these positive features, the incrementalist approach also highlights an important idea about normative theorising in general, which has not been discussed in the article up to this point and, while it is not necessarily controversial, it merits a strong emphasis here. The idea which I have in mind is that normative principles should not be thought of as universally applicable for any configuration of facts about the world, but rather that particular normative principles should be applied under particular configurations of facts about the world. This does not imply that any change in the set of facts about the world should lead us to adopt new normative principles. What it does imply, however, is that, regardless of whether the principles are generated at the level of ideal theory or at the level of non-ideal theory, their implementation is dependent on the actual state of the world, since the first category of principles might be much more undesirable/unfeasible under more fact-sensitive assumptions and the first category of principles might be less desirable than others if some soft constraints have changed the state of the world in which they were originally implemented.
normative incrementalism which I have outlined in this article should prove to be a fertile
ground for the creation of systematic approaches to political philosophy, that can at the same
time be action-guiding for the current world and move us in the correct direction in the long-
run.

Bibliographical references:
decision", Theoria, 70(1), pp.3-21.
Arneson, R. (2000a), "Welfare should be the currency of justice", Canadian Journal of
Philosophy, 30(4), pp.497-524.
Brown, C. (2005), "Priority or Sufficiency...or Both?", Economics and Philosophy, 21, pp. 199-220.


Uberti, L. J. (2014), "Good and Bad Idealizations in Political Theory", Theoria, 80, 205-231.