Version 5.0 // Oct 2014

Against Ideal Guidance

David Wiens

Abstract. The prevailing wisdom among political philosophers claims that political ideals provide normative guidance for unjust and otherwise nonideal circumstances. This paper has two objectives. The first is to develop a model of the logical relationship of moral evaluative considerations to feasibility considerations in the justification of normative political principles. The second is to use this model to demonstrate that political ideals are uninformative for the task of specifying the normative principles we should aim to satisfy amidst unjust or otherwise nonideal circumstances. The argument implies that social scientists have an essential contribution to make to the normative theoretical enterprise.

Analyzing political ideals is a core business of contemporary political philosophy.[1] This enterprise typically abstracts from many nonideal features of our world to better view the constitutive features of a fully just society. To some, this is a dubious and mystifying practice. Amidst the pressing injustices all around us, we urgently need to know how to regulate our political affairs in manifestly nonideal circumstances; normative principles that characterize a fully just society seem unfit to provide the required guidance. Political ideals are neither necessary nor sufficient for comparing feasible institutional reforms with respect to their potential for advancing justice (Sen, 2009). In addition, the application of political ideals to real world circumstances is alleged to be "misguided" and "inappropriate" at best (Farrelly, 2007; Wiens, 2012), dangerously "ideological" at worst (Geuss, 2008; Mills, 2005). So the skeptics say.

These skeptical arguments suffer serious flaws. Sen's criticisms mischaracterize the role of ideals in the Rawlsian methodological paradigm he rejects (Valentini, 2011; Gilabert, 2012*a*). Moreover, proponents of the prevailing practice readily concede that

Author's note. Thanks to Dave Estlund, Holly Lawford-Smith, Shanna Slank, Nic Southwood, Lachlan Umbers, and anonymous referees for helpful suggestions. The discussion in section 3 has benefited greatly from numerous conversations with Geoff Brennan. Support from the Australian Research Council (Discovery Grant DP120101507) is gratefully acknowledged.

1 For lack of a better term, I use "political ideal" ("ideal" for short) to denote a set of normative principles that specifies, in general, certain constitutive features of a fully just state of affairs (I will refine this throughout the paper). To avoid confusion, I refer to abstract moral and social values such as liberty and equality as "(basic) evaluative criteria" ("values" for short). For some, this distinction might recall Hamlin and Stemplowska's (2012) distinction between "ideal theory" and "theory of ideals". I enumerate important differences toward the end of section 3.

Penultimate version. To appear in *Journal of Politics*

ideal theoretic principles do not straightforwardly apply to nonideal circumstances. Instead, ideals proffer normative targets — principles that we should aim to satisfy in our world, if only by incremental steps (Robeyns, 2008; Simmons, 2010; Valentini, 2009). Political ideals can also help settle the criteria we use when evaluating intermediate reform options (Gilabert, 2012*a*; Jubb, 2012; Sangiovanni, 2008). Either way, a political ideal remains a crucial input for the specification of morally progressive principles for nonideal circumstances.

This paper makes two contributions to recent discussion concerning the role of political ideals in the specification of normative principles for nonideal circumstances. First, I sketch a general model to distinguish three core components of typical normative political theories and clarify their logical interrelationships. Briefly, these components are: *evaluative criteria*, which buttress a comparative ranking of possibilities; *empirical constraints*, which delimit the possibilities for jointly realizing the chosen evaluative criteria; and *directive principles*, which specify the possibilities we are to realize by demarcating the lines between obligatory, permissible, and prohibited interpersonal conduct or institutional schemes.[2] This model offers clarity and focus to a debate that risks disorder due to equivocal usage of terms like "ideal theory", "normative principle", and "practical guidance". Second, I use this model to improve upon existing skeptical arguments, demonstrating that political ideals contribute nothing to our reasoning about what to do amidst nonideal circumstances. Political ideals are not simply unnecessary and insufficient for specifying the directive principles we should aim to satisfy in our world; they are utterly uninformative for this purpose.

My argument implies a need for deeper partnership between political philosophers and social scientists in the normative theoretical enterprise. The traditional division of labor has political philosophers specify directive principles before turning to social scientists to sort out the implementation issues. My argument implies that this division is mistaken (cf. Miller, 2008). The normative theoretical role of social science is not limited to working out implementation issues; social scientists serve an integral role in specifying the directive principles we should aim to satisfy in our world.

Let me register three caveats. First, I have no ambition to propose a "correct" characterization of ideal theory (for recent surveys, see Hamlin and Stemplowska, 2012; Valentini, 2012). When convenience dictates, I use the labels "ideal theory" and "nonideal theory" stipulatively, as follows: an *ideal theory* presents a specification and analysis of

2 Importantly, directive principles are *moral* principles rather than what might be called "all-things-considered" principles; they are meant to specify permissions and obligations *from the standpoint of moral theory*, not simply from a policy standpoint.

a political ideal (with the concept of political ideal being progressively refined over the course of the paper); a *nonideal theory* presents a specification of directive principles for unjust circumstances. I aim to show that the conventional practice of specifying and analyzing political ideals fails to guide nonideal theorizing in any of the ways "guidance" is typically conceived. Second, I say nothing here about how best to do nonideal theory; discussion of that topic is beyond the scope of this article. Instead, my argument concerns whether political ideals should serve as reference points for nonideal theory. I conclude that they should not — however we should do nonideal theory, it should not be with an eye to political ideals. Third, I do not argue that political ideals are useless *for all purposes* (as other skeptics claim); for example, exploration of political ideals might help inspire agents to pursue feasible justice-enhancing reforms (cf. Gheaus, 2013). For all I say here, we may yet have some reason to investigate political ideals. I simply argue that the need to identify directive principles for nonideal circumstances like ours provides no such reason.

In section 1, I briefly trace the key threads of the debate over the normative theoretical role of political ideals. In section 2, I use Rawls's and Nozick's theories of justice to uncover key insights regarding the justification of political ideals. I generalize these insights in section 3, developing a general model of normative political theories. This model clarifies the relationship between basic moral values and empirical constraints with respect to the justification of directive principles. I then use this model to press two objections against the claim that political ideals provide guidance for nonideal theory. In section 4, I show that analyzing political ideals plays no role in identifying the directive principles we should aim to satisfy in nonideal circumstances. In section 5, I show that political ideals are unsuitable for use as evaluative standards when comparing feasible institutional alternatives. Section 6 concludes.

1. IDEAL GUIDANCE AND ITS CRITICS

For now, let's say that a *political ideal* is a set of directive principles that specifies, in general, certain constitutive features of a fully just state of affairs (this is somewhat crude, but will be refined as we go along). The prevailing approach to normative political theory starts by analyzing political ideals because, following Rawls, "it provides... the only basis for the systematic grasp of [the] more pressing problems" that arise in unjust or otherwise nonideal circumstances (Rawls 1999, p. 8; cf. Simmons 2010, p. 34 and Valentini 2009, p. 333). "Ideal theory guides nonideal theory" is an apt slogan for the prevailing wisdom. This "ideal guidance view" (Wiens, 2012) is not without its critics. The core complaint is that political ideals are ill-suited to offer normative guidance for a nonideal world like

ours. Ideal analysis typically assumes an idealized, perhaps infeasible, world. As a result, political ideals abstract from urgent normative issues that arise from (e.g.) acute resource scarcity (Farrelly, 2007), racial or gender injustice (Mills, 2005), or profound political conflict (Geuss, 2008; Williams, 2005). How can normative theories that abstract from real world challenges help us figure out what to do in a markedly nonideal world?

Among skeptics, Amartya Sen (2009) has arguably attracted the most attention. Sen attributes several problems to ideal theory ("transcendental institutionalism" as he calls it), but his claim that an analysis of perfectly just institutions is neither necessary nor sufficient for identifying ways to advance justice has made the biggest splash. A political ideal is not *necessary* because we can identify which of two alternatives ranks higher according to some metric without identifying which alternative ranks highest. To take Sen's example, we need not know that Mount Everest is the highest mountain to determine that Mount McKinley is higher than Kilimanjaro (Sen, 2009, 101f). A political ideal is not *sufficient* because our world deviates from the political ideal in many ways and there are many metrics we might use to compare these deviations. Simply knowing what perfect justice demands is not enough to carry out this comparative exercise (Sen, 2009, 98ff). Hence, we should focus less on analyzing political ideals and invest more in comparative evaluation of feasible options for advancing justice in a nonideal world.

Sen's argument faces several challenges, most of which remain unanswered. First, a political ideal is not meant to identify a uniquely just *institutional scheme*; rather, an ideal proffers a set of *general principles*, which can be fulfilled by multiple institutional schemes (Valentini, 2011; Gilabert, 2012*a*). Second, if the aim of nonideal theory is to identify transitional paths toward a perfectly just society (as many ideal guidance proponents suppose), then we need to identify the pinnacle of justice to determine which lesser peaks lay along the best transitional path. Sen's mountain analogy neglects this point: if my aim is to reach the *highest* peak (full justice) and not simply a higher one (more justice), then (in our world) I need to identify Mount Everest (Simmons, 2010, 34f).[3] Third, ideal guidance proponents concede that analyzing political ideals cannot settle every issue that arises in nonideal normative analysis. Yet, ideal analysis can help settle the relevant criteria for comparing and evaluating feasible options (Gilabert, 2012*a*; Sangiovanni, 2008; Swift, 2008).[4] Finally, the claim that political ideals are useless for

3 This challenge assumes that we should aim to reach full justice (the highest peak). Schmidtz (2011) offers a reply on Sen's behalf: we shouldn't aim for full justice because there's no such thing. So we needn't know anything about full justice. Although I'm sympathetic with Schmidtz's point, I present a different response to Simmons's challenge in section 4 below: namely, that we cannot justify a reasonable expectation that a political ideal presents an appropriate target for nonideal theory.

4 My argument in section 5 addresses this claim.

nonideal theory does not follow from the fact that they are neither necessary nor sufficient for specifying nonideal directive principles. A travel guide for Argentina or Zambia can prove useful despite being neither necessary nor sufficient for having an enjoyable travel experience. A travel guide ceases to be useful when it is misleading or uninformative. Similarly, analyzing political ideals might enhance our nonideal normative theorizing despite being neither necessary nor sufficient for specifying nonideal directive principles. Thus, ideal theory skeptics must go beyond Sen's arguments to show that political ideals are *misleading* or *uninformative* when specifying the directive principles we should aim to (eventually) fulfill in nonideal circumstances. Showing this is the task of sections 4 and 5. The next two sections develop a general model of normative political theory to regiment the terms of debate and facilitate exposition of my skeptical arguments later on.

2. POLITICAL IDEALS: TWO EXEMPLARS

Here's a rough sketch of the structure I wish to build. In general, we evaluate states of affairs according to the extent to which they realize certain basic moral and social values. We ultimately care about, for instance, the extent to which people are free and physically secure, the extent to which their moral equality is respected, or the extent to which communities accommodate diverse lifestyles. However, the practical implications of our commitment to these values are often vague. This is where directive principles do their work: they codify these vague implications. A set of directive principles, P, characterizes a state of affairs, s, by codifying the normatively constitutive features of s. Hence, P comprises the general rules that obtain at s to specify agents' claims vis-a-vis each other and to resolve potential conflicts among them; for example, P might enumerate the scheme of rights and duties that obtains at s. Since states of affairs can be identified by the extent to which they realize certain basic values, P characterizes the balance among basic values realized by the normatively constitutive features that obtain at s. If P characterizes s in this sense, then P *reflects* certain basic values to the extent that s *realizes* those values.[5] P is preferable to another set of principles, P', if P reflects the specified basic values to a greater extent than P' does. To wit, suppose we could successfully implement P or P'. If satisfying P leads to people generally living freer and

5 With some abuse of language, "reflect" and "realize" are meant to be neutral among several relations of potential interest here. For instance, we might care about certain causal relations: that a state of affairs or a principle *bring about*, *promote*, or *enhance* material equality. Or we might care about certain noncausal relations: that a state of affairs or a principle *honor* or *respect* individuals' moral equality, or *express*, *convey*, or *demonstrate* equal concern for all. Hereafter, I will speak of principles *reflecting* basic values and states of affairs *realizing* basic values.

more secure lives, better respects individuals' moral equality, or better accommodates diversity than implementing P' does, then we say that P better reflects the specified basic values than P' does. Plausibly, P is morally preferable to P' in virtue of this fact. Given all this, let's now say that a *political ideal* is a set of directive principles that best reflects an ideal balance of basic moral and social values, that is, the balance of basic values realized at a fully just state of affairs.

In the next section, I regiment the core components of this sketch by presenting a general model of normative theorizing wherein *directive principles* are taken to be justified in light of the extent to which they reflect certain *basic evaluative criteria* given a set of *empirical constraints*, which consist of certain assumptions about which states of affairs can be realized.[6] Developing this model requires refinement along three dimensions: directive principles, evaluative criteria, and empirical constraints. I start, in this section, by appeal to examples.

Rawls famously argues for two principles of justice (three, depending on how one counts). The first states that each member of a society has an equal claim to the most extensive set of basic liberties compatible with everyone enjoying a similar set (the "equal basic liberties principle"); the second states that socioeconomic inequalities are permitted only insofar as they maximize the prospects of the least advantaged members (the "difference principle"), subject to the requirement that all members have effectively equal opportunities to occupy the social positions to which socioeconomic shares are to be attached (the "fair equality of opportunity principle" respectively) (see Rawls, 1999, p. 266; bare page numbers refer to this work for the following discussion). Ideally, these principles are to be "lexically ordered" with respect to each other: we are to ensure that members have equal basic liberties first, then ensure that members enjoy equal opportunities, before finally moving on to address members' socioeconomic prospects (p. 266). Thus, ideally, individual liberties, equality of opportunity, and socioeconomic gains cannot be traded off against each other (sec. 11). In the preceding terms, this set of principles — the two principles of justice and the priority rules — codifies, in general, the normatively constitutive features of a fully just institutional scheme (secs. 1–2). They are directive

6 It is important to understand at the outset that the proposed model concerns the structure of the logic whereby directive principles are justified, not the actual processes by which the content of directive principles is discovered. I do not claim that evaluative criteria are *epistemically* prior to directive principles, that is, that we first articulate a complete set of evaluative criteria and hold these fixed in our subsequent specification of directive principles. Rather, the upshot of the model is that evaluative criteria are *logically* prior to directive principles — optimal realization of certain evaluative criteria *explains why* certain directive principles are taken to be justified. I address this point further below. (Thanks to an anonymous reviewer for pressing me to clarify this point.)

principles (as I understand that concept): they serve to delimit the sets of obligatory, permitted, and prohibited institutional schemes.

What's left to show is that Rawls's principles are meant to characterize an *ideal balance of certain basic values*, that is, the balance of basic values realized by a fully just institutional scheme. To see that this is so, consider the role of the original position in Rawls's argument. The original position represents a hypothetical choice situation, which is constructed to model certain moral constraints on the selection of directive principles of justice: "the original position. . . [is] an expository device which sums up the meaning of these conditions and helps us to extract their consequences" (p. 19; sec. 4 *passim*). "[T]hese conditions" refers to "the restrictions it seems reasonable to impose on arguments for principles of justice, and therefore on these principles themselves" (p. 16; see also p. 19 and sec. 20). What are these conditions? In setting out his guiding intuitions, Rawls indicates that principles of justice should sustain "a system of cooperation designed to advance the good of [participants]", one that coordinates "the plans of individuals. . . so that their activities are compatible with one another" and is stable over time (pp. 4–6). Further, principles of justice should not be tailored to any individual's particular circumstances or serve any particular interests (p. 16). The veil of ignorance is introduced to situate parties fairly with respect to each other, to ensure that the chosen principles eschew "arbitrary distinctions. . . between persons in the assigning of basic rights and duties", thereby respecting individuals' freedom and fundamental moral equality (p. 5; cf. pp. 17, 104, sec. 24). These moral conditions modeled by the original position operationalize certain "commonly shared" basic values (p. 16) and, thus, represent *basic evaluative criteria* (as I understand that concept).[7] Rawls's conception of justice is then selected by parties in the original position from among a menu of sets of principles, where each set is conceived as a proposal for codifying the practical demands of our evaluative criteria on institutional schemes.[8] These alternatives are then "ranked by their acceptability to persons so circumstanced", that is, to persons situated within the original position (p. 16; cf. sec. 21). Sets of directive principles are thus comparatively evaluated by parties in the original position according to the extent to which their institutional instantiations realize the evaluative criteria modeled by the original position (cf. Pogge, 1989, pp. 36–47).

This comparative evaluation is done under Rawls's assumptions about the empirical constraints delimiting feasible states of affairs, which identify the ways in which his basic

7 See also Rawls (2001, §6).

8 The sets of principles included on the menu are themselves subject to certain formal conditions, such as generality, universality, and publicity (sec. 23) These conditions are plausibly viewed as modeling a certain kind of fairness or impartiality, which qualify as evaluative criteria in my sense.

evaluative criteria can be jointly realized. For example, material resources are assumed to be moderately scarce, though not so scarce as to block mutually beneficial social cooperation (p. 110). Members of society are also assumed to have a "sense of justice", which Rawls defines as "an effective desire to comply with the existing rules and to give one another that to which they are entitled" (p. 274f). Famously, interactions with other societies are not assumed to constrain the realization of the basic values modeled by the original position (p. 7).

Summarizing thus far, we've seen that Rawls's original position models, in a more or less precise way, (1) a set of intuitively plausible moral conditions on the specification of principles of justice, and (2) a set of empirical constraints on the joint realization of those moral conditions. In my terms, the first set represents Rawls's *basic evaluative criteria* and the second set represents Rawls's assumed *empirical constraints*. The remaining issue concerns the relationship of these components to Rawls's principles of justice. Rawls settles the issue in no uncertain terms: a set of directive principles is justified if it would be selected from among a menu of alternatives by parties subject to the constraints modeled by the original position: "the question of justification is settled by working out a problem of deliberation: we have to ascertain which principles it would be rational to adopt given the contractual situation" (p. 16). Nothing turns on the fact of hypothetical agreement though. The original position device is simply used to assess alternative sets of principles according to their consistency with the conditions modeled by the choice situation (p. 19). Rawls goes on to argue that the original position — his particular construction of the contractual situation — is the "most favored" interpretation because it "embodie[s]" the moral criteria "we do in fact accept" or "can be persuaded [to accept] by philosophical reflection" (p. 19).[9] Put simply, the original position is the best model of widely shared evaluative criteria (among members of a liberal democratic society). Rawls then aims to show that his set of principles "is the only choice consistent with the full description of the original position" (p. 104). Thus — and this is the key point — Rawls understands his set of directive principles to be justified in virtue of the fact that it best reflects — is most consistent with — certain basic evaluative criteria given the assumed empirical constraints. Put differently, Rawls's principles are meant to codify the normatively constitutive features of a fully just institutional scheme, understood as a member of the set of institutional schemes that optimally realize certain basic values

9 There is a complication here introduced by the method of reflective equilibrium, which requires that "we work from both ends" in constructing the original position, so that the principles it yields are both consistent with the moral constraints modeled and cohere best with our considered judgments in particular cases (see pp. 18–19, 105). I deal with this complication below. (Thanks to an anonymous reviewer for drawing my attention to this complication.)

given the institutional possibilities consistent with the assumed empirical constraints.[10]

Let's briefly consider a second example. Nozick argues for the following set of directive principles: a side-constraint prohibiting physical aggression against another person (Nozick, 1974, pp. 28–35); and an "entitlement theory" of distributive justice, which consists of principles regulating appropriation and transfer of property holdings and a principle of rectification to address unjust appropriations or transfers (Nozick, 1974, pp. 150–160; bare page numbers refer to this work for the following paragraphs). Importantly, Nozick's distributive principles are "historical" — they attend solely to the character of the interpersonal transactions that bring about a particular distribution and not to any structural features of the distribution itself — and "unpatterned" — they do not require that holdings be distributed according to any general criterion (such as need, virtue, marginal product, and so on) except that of voluntary choice. Unlike Rawls, Nozick doesn't derive his principles from a model that systematizes his evaluative criteria or assumed empirical constraints. Yet, viewed from a more general perspective, Nozick proceeds in much the same way as Rawls.

Consider first Nozick's argument for a side-constraint against physical aggression. Rights are to be viewed as side-constraints rather than in a consequentialist manner because side-constraints better respect "the inviolability of others", which is supposed to encapsulate the Kantian dictum that people be treated as ends and never solely as means to achieving others' goals (pp. 30–32). In addition, side-constraints better respect the "separateness of persons" than a consequentialist view of rights (pp. 32–33). Finally, side-constraints are argued to better reflect an interest in being self-directing and exercising our capacity to create meaning (pp. 42–45, 50). At bottom, each of these criteria is meant to capture an overarching interest in autonomous personal agency. So, for Nozick, a side-constraint against physical aggression is justified because it best reflects his specified basic evaluative criteria.

Notably, Nozick's argument for viewing rights as side-constraints is framed by a set of assumptions about how the world works, summarized by the core features of a Lockean state of nature: a world in which people "generally satisfy moral constraints and generally act as they ought" (p. 5) but are prone to "overestimate the amount of harm or damage they have suffered", which leads them "to attempt to punish others more than proportionately and to exact excessive compensation" (p. 11). Hence, we transition from the state of nature to political society to avoid the deficiencies of decentralized rights enforcement. Alternative sets of directive principles for regulating this transition are thus considered in

10 See Rawls (2001, part 1, *passim*) for corroboration of the interpretation given here.

light of the extent to which they reflect the value of autonomous agency given a broadly Lockean state of nature. A side-constraint against aggression emerges as a justified limit on permissible transition steps because, among the alternatives, it best reflects Nozick's evaluative criteria given the assumed empirical constraints. (Would side-constraints best reflect these criteria if the state of nature were closer to Hobbes's depiction? Would some uses of force for the sake of promoting social cooperation better reflect our interest in autonomous agency in a chaotic "war of all against all"? Cf. p. 5.)

Let's now turn to Nozick's argument for the entitlement theory. Principles of distributive justice must be historical and unpatterned because such principles better reflect the specified evaluative criteria than end-state or (historical but) patterned principles do — the values of free and voluntary choice; the inviolability and separateness of persons; the capacity to create meaning (pp. 159–160, 167, 171–172). Nozick's most forceful argument here is his famous Wilt Chamberlain case. Nozick imagines a world where Wilt Chamberlain signs a contract that entitles him to twenty-five cents from each ticket sold to watch him play basketball. Accordingly, a box labelled "Wilt Chamberlain" is set at the arena's gates and each spectator willingly drops a quarter into the box as she enters the gate. Nozick's point is this: whatever the distribution we start off with at the beginning of the basketball season (pick your favorite one), the distribution that results from a season's worth of voluntary payments to Wilt will deviate from the initial distribution. Hence, "no end-state principle or distributional patterned principle of justice can be continuously realized without continuous interference with people's lives" (p. 163). According to Nozick, such interference is incompatible with autonomous personal agency.

The justification of Nozick's entitlement theory, too, depends on certain assumptions about the world. To name but a few: that the price for watching Wilt is separable from the price for watching the other players; that spectators regard as fair the price negotiated by the team's owners on their behalf; that the price is presumed to be fair without regard for the factors determining the parties' relative bargaining strength. Alternative sets of distributive principles are thus considered in light of the extent to which they reflect the value of autonomous agency assuming (at a minimum) a world without transaction costs or illegitimate bargaining advantages. The entitlement theory is taken to be justified because historical and unpatterned distributive principles best reflect Nozick's evaluative criteria given the possibilities consistent with these assumed circumstances. (What if the price for watching Wilt is not separable from the price of admission and we are all willing to pay the one dollar admission price but unwilling for twenty-five cents of that dollar to go to Wilt, preferring for more of the admission price to go to the other players? Has each of us voluntarily transferred twenty-five cents to Wilt in this case? Or what if

Wilt is able to extract such a high price for his services because he secretly bribes his chief rival, Bill Russell, to play overseas, thus allowing Wilt to corner the market for star basketball players? More generally, would we find it plausible that the entitlement theory best reflects the basic value of personal liberty given certain market failures that interfere with our ability to direct our holdings as we choose?[11])

3. NORMATIVE POLITICAL THEORY: A GENERAL MODEL

Rawls's and Nozick's theories are exemplars of the general model I aim to develop, wherein a set of directive principles is justified in virtue of the fact that it optimally reflects certain basic evaluative criteria given a set of empirical constraints. Let me now sketch that model formally, to facilitate an exploration of its implications for the ideal guidance view.[12]

We start with the set of all logical possibilities, denoted W. For expository purposes, we represent possibilities with possible worlds, each of which models a complete way the world might be.[13] The key function of the possible worlds technology here is to facilitate comparison of counterfactual states of affairs (Lewis, 1973). Morally speaking, we evaluate possible worlds according to a set of *evaluative criteria*, which we denote $V = \{v_1, \ldots, v_n\}$. Each element of this set presents an analysis or a representation of a basic value. Basic moral and social values are things like freedom, equality, security, peace, community, and so on. A set of basic evaluative criteria identifies the values to which our comparative assessments attend and specifies their content. For example, the evaluative criterion associated with the value of freedom specifies whether to conceive of freedom in terms of capabilities, non-domination, absence of external impediments to action, self-governance, and so on. The criterion associated with the value of equality answers

11 Two anonymous reviewers have objected that the point of the Wilt Chamberlain case is narrower than I've indicated here — namely, that Nozick only aims to demonstrate that maintaining a patterned distribution requires continuous interference with people's choices. In reply, it seems clear to me that Nozick's rhetoric here strongly suggests (if not implies) something wider than this narrow point. The relevant section is entitled "How *liberty* upsets patterns" (my emphasis), and Nozick's rhetoric throughout indicates that the interference required to maintain a patterned distribution is meant to be viewed as incompatible with autonomous personal agency. It is this wider point — that historical and unpatterned distributive principles best reflect the value of autonomous agency — that I take to depend on the enumerated empirical assumptions. The questions with which I end this section are meant to indicate that we may no longer find certain kinds of interference with market transactions incompatible with autonomous agency in a world with transaction costs or illegitimately gained bargaining advantages.

12 I stress that my presentation here is only a sketch; I must leave investigation of certain technical niceties for another paper.

13 Alternatively, we might restrict our attention to the way some salient part of worlds, or to social structures, or to some other specification of possibilities.

the "equality of what?" and "what's the point?" questions (among others) — Should we equalize resources, capabilities, or opportunities? Should egalitarians seek to eliminate the influence of bad luck on socioeconomic distributions or undermine social domination and status hierarchies? Rawls's representation of moral equality by the veil of ignorance or Nozick's explication of autonomy in terms of the inviolability and separateness of persons are further examples.[14]

We treat evaluative criteria as variables that can take on any number of distinct structures: they might be continuously realizable (i.e., realizable along an infinitely graded scale), dichotomous (i.e., realized or not), or categorical (i.e., a finite number of discrete gradations). For the purposes of moral evaluation, we identify possible worlds by the values assigned to these evaluative variables. Hence, we can depict each possible world as a vector, $w = (v_1, \ldots, v_n)$, thereby locating each world somewhere within the N-dimensional space defined by our evaluative criteria. We can construct a ranking among possible worlds via a series of comparative evaluative judgments, which embodies the fact that we regard some worlds as better realizing our evaluative criteria than others. For our purposes, this ranking can have any structure we like, so long as it can be represented by a binary relation, here denoted $\succeq_V$, that is reflexive, transitive, and complete.[15] We interpret $\succeq_V$ as follows: for all worlds w and w' in W, $w \succeq_V w'$ if and only if w is at least as morally desirable as w' in light of the specified evaluative criteria. This $\succeq_V$ relation yields an indifference map for the space defined by the set of evaluative criteria.

As noted in the last section, normative political theories assume a set of *empirical constraints* on the extent to which our evaluative criteria can be jointly realized; we denote this set $C = \{c_1, \ldots, c_n\}$. These constraints circumscribe the set of possible worlds that are not ruled out by C, denoted $W(C)$; that is, $W(C)$ is the set of worlds at which each $c_i \in C$ is satisfied. For example, Rawls's assumptions rule out worlds at which political communities are interdependent and people are "ready to act unjustly should doing so promise some personal advantage" (Rawls, 1999, p. 498); Nozick's assumptions rule out worlds at which the state of nature is as Hobbes depicted and certain market

14 Compare Swift's view, according to which "context-independent philosophy" fills two roles: "formal or conceptual analysis yielding precision about the various values at stake, how they relate to one another, and so on"; and offering "substantive or evaluative judgments about the relative importance or value of the different values at stake" (Swift, 2008, p. 369). Swift's use of "values" apparently conforms to my use of "evaluative criteria". Importantly, he does not seem to hold that directive principles (e.g., Rawls's two principles) serve this evaluative purpose (Swift, 2008, p. 382). This contrasts with those who argue that ideal directive principles serve an evaluative function. I consider this latter claim in section 5.

15 Reflexive: $w \succeq_V w$ for all $w \in W$; transitive: if $t \succeq_V u$ and $u \succeq_V w$, then $t \succeq_V w$, for all $t, u, w \in W$; complete: $u \succeq_V w$ or $w \succeq_V u$ for all $u, w \in W$. We can relax transitivity or completeness as necessary; I assume these conditions to simplify the exposition.

failures are present. The point here, to be clear, is not that these assumptions rule out the possibility of realizing the specified evaluative criteria in worlds that violate these constraints; instead, the point is that comparative assessment of the extent to which alternative sets of principles reflect the specified evaluative criteria is limited to the set of possibilities that satisfy the assumed constraints.[16]

Recall that a political ideal is a set of directive principles that optimally reflects certain basic moral and social values given a set of assumed empirical constraints. Let's explicate this idea more precisely now. Let $B(C)$ denote the set of morally best, or optimal, possible worlds given C, the set of constraints. $B(C)$ is the subset of worlds in $W(C)$ for which there is no world in $W(C)$ that is more morally desirable.[17] Refining our earlier analysis, we now say that a set of *directive principles*, $P = \{p_1, \ldots, p_n\}$, characterizes the deontic implications of certain evaluative criteria by enumerating the normatively constitutive features of $B(C)$. Loosely following the standard analysis of deontic expressions like "ought" and "may" (e.g., Kratzer 1991; see Charlow forthcoming for helpful discussion), we might unpack this as follows. Let ϕ denote some general kind of institutional scheme or interpersonal conduct; then:

- ϕ is *permitted* if (and only if) there is at least one world in $B(C)$ at which ϕ is realized;
- ϕ is *prohibited* if (and only if) ϕ is not permitted (i.e., there is no world in $B(C)$ at which ϕ is realized);
- ϕ is *obligatory* (or *required*) if (and only if) $\neg\phi$ is prohibited (i.e., ϕ is realized at every world in $B(C)$).

For example, given Rawls's specification of C, Rawls's difference principle is recommendable as a deontic requirement of justice if every world in $B(C)$ realizes an institutional scheme that satisfies the difference principle. More generally, a set of directive principles optimally reflects certain evaluative criteria when it enumerates which kinds of institutional scheme or interpersonal conduct are permitted, prohibited, or obligatory in view of $B(C)$. On the model developed here — and exemplified above by Rawls and Nozick — a set of directive principles, P, is justified in virtue of the fact that P codifies the normatively constitutive features of the set of morally optimal worlds, $B(C)$. If P characterizes $B(C)$ in this way, then we say that P optimally reflects the specified basic moral and social values in light of the assumed empirical constraints.

16 Thanks to an anonymous reviewer for alerting me to this potential confusion.

17 Precisely: $B(C) = \{w \in W(C) : w \succeq_V u \text{ for all } u \in W(C)\}$.

There are two points worth noting about the model developed here. The first is that it is compatible with nonconsequentialist political theories — it can accommodate sets of directive principles that do not have a maximizing, goal-oriented structure without distorting their distinctively nonconsequentialist structural features. This is a primary reason for starting with Rawls's and Nozick's theories, both of whom are avowed opponents of consequentialist normative theories. At bottom here is a distinction between two kinds of goals. The first kind comprises the moral goals held out for a group of people by a set of directive principles (e.g., optimize welfare, or equality, or liberty, or some combination of the three); the second kind comprises the goals of normative theorists, namely, to identify the directive principles that best reflect our moral evaluative criteria. Given this distinction, a concern to specify directive principles that optimally reflect selected evaluative criteria need not result in principles that advise moral agents to maximize the realization of some moral goal. For example, the directive principles that optimally reflect respect for individual autonomy need not prescribe actions or institutions that minimize the number of rights violations; they might, instead, prescribe a set of constraints that prohibits certain kinds of actions whatever the consequences (cf. Nozick, 1974, pp. 28–30).

The second important point is that the model need not presuppose that we have settled our evaluative ranking prior to specifying our directive principles. For Rawls, "we work from both ends" to specify the original position, seeking a "reflective equilibrium" between our judgments about the hypothetical contractual situation that best reflects our evaluative criteria and the fit of the resultant principles with our considered judgments about particular cases (Rawls, 1999, pp. 18–19, 40–46, 105). In the terms of the model, the method of reflective equilibrium implies that we occasionally revise our evaluative ranking to ensure that the resultant directive principles supply the best fit with our considered intuitions about particular cases. This poses no problem for the model. When the directive principles derived from an initial specification of our evaluative ranking are at odds with our considered convictions, this just means that our initial specification doesn't adequately capture the evaluative ranking latent in our intuitions about cases. In response, we might revise the evaluative criteria to better fit the comparative ranking implied by our considered convictions. No doubt our exposition of certain considered convictions might well be presented in terms of directive principles; for example, that no justified set of directive principles can include a permission to keep slaves. But this is readily interpreted as revealing the (latent) comparative conviction that any world at which institutions tolerate chattel slavery ranks below some set of feasible worlds at which institutions occlude slavery. A set of evaluative criteria is justified by the method of

reflective equilibrium if it coheres with our considered *evaluative* convictions as revealed by intuitive judgments about particular cases.[18] The point of the model is to show that evaluative criteria precede directive principles in the order of *justification*, not necessarily in the order of discovery. A specification of evaluative criteria is justified together with the comparative ranking of possibilities it supports in reflective equilibrium; the justification of a set of directive principles is explained by the fact that it reflects an optimal balance of certain evaluative criteria.

The model presented here might recall that developed by Hamlin and Stemplowska in characterizing their distinction between ideal theory and a "theory of ideals". A theory of ideals specifies the criteria we use to evaluate alternative institutional schemes, while ideal theory specifies institutional arrangements that realize the specified evaluative criteria subject to feasibility constraints (Hamlin and Stemplowska, 2012, p. 53). But their model differs from mine in two respects worth noting here. First, they claim — following a suggestion from G.A. Cohen (2003, pp. 244f) — that ideal theory is primarily concerned with institutional design, whereas I maintain that ideal theory is primarily concerned with specifying general directive principles, each set of which might be instantiated by multiple institutional schemes. Second, they identify Rawls's equal basic liberties principle as an evaluative criterion in their sense, whereas I have presented that principle (together with Rawls's other principles) as directive principles that are meant to reflect more basic evaluative criteria. Hamlin and Stemplowska's model sensibly indicates that we should try to implement the institutional scheme that best satisfies the appropriate directive principles. But this obscures the deeper issue of how we specify appropriate directive principles in the first place. Given my model, directive principles are not basic in the way Hamlin and Stemplowska's model suggests; they are meant to reflect more basic evaluative criteria. Further, *pace* Hamlin and Stemplowska's claim to the contrary (2012, p. 55), I have shown that Rawls's (and others') specification of his favored directive principles does account for certain feasibility constraints; it simply does so by assumption rather than careful investigation of the actual world.

Hamlin and Stemplowska's treatment of Rawls's principles of justice as criteria for evaluating alternative institutional schemes reveals an important ambiguity in the existing literature on ideal theory, one I have sidestepped to this point. Thus far, I have treated political ideals as serving a *directive* role (others might say "prescriptive" or "practical"); I have presented Rawls's and Nozick's principles of justice as demarcating the lines between obligatory, permissible, and impermissible actions or institutional schemes. But, following Hamlin and Stemplowska (among others), we might think that political ideals

18 Thanks to an anonymous reviewer for pressing me to clarify this point.

best serve an *evaluative* role; we might treat Rawls's or Nozick's principles as implying something about the ranking of possibilities from a moral standpoint, as indicating which possibilities are better than others in view of the salient moral considerations.

This ambiguity is transposed to discussions of how ideal theory provides guidance for nonideal theory. Ideal theory might be taken to guide nonideal theory in a directive sense or an evaluative sense. The ideal guidance view thus splits into two distinct interpretations. On the *target view*, political ideals are understood directively, as specifying a set of general principles that we should aim to satisfy, even if only approximately. This interpretation is bolstered by a handful of plausible metaphors — a political ideal as a compass or a map (see, among others, Gilabert, 2012*b*; Robeyns, 2008; Simmons, 2010; Valentini, 2009). In contrast, the *benchmark view* conceives of political ideals evaluatively, as specifying a standard by which we rank feasible worlds. On this interpretation, a political ideal does not straightforwardly specify the directive principles we should aim to satisfy, but instead helps us comparatively evaluate alternative sets of directive principles in light of our feasible options (see Gilabert, 2012*a*; Jubb, 2012; Sangiovanni, 2008; Stemplowska, 2008). Rejecting the ideal guidance view requires rejecting both its target and benchmark interpretations. The model developed in this section provides a basis for both tasks.[19] I argue against the target view in the next section and against the benchmark view in section 5.

4. AGAINST IDEAL TARGETS

In this section, I argue that political ideals specified with little regard for feasibility considerations do not provide an appropriate target for real world reform efforts. For expository purposes, I present my argument in terms of Rawls's theory, although the argument readily generalizes, given the model developed in the last section.

Let $W(F)$ denote the set of feasible worlds, the set of worlds that are consistent with an accurate specification of the feasibility constraints that obtain in the actual world.[20] Let $B(F)$ denote the set of morally optimal feasible worlds — the subset of worlds in $W(F)$

19 One might think that my presentation of the model is prejudiced against the benchmark view. This is not so, as I will make clear in section 5. The issue raised by the benchmark view is whether the (directive) principles that characterize a fully just state of affairs can serve some sort of evaluative purpose with respect to the set of feasible options.

20 As will become apparent, the argument does not depend on any particular analysis of the feasible set, save to say that the feasible set should include outcomes that may not be immediately realizable but can be realized at some point in the foreseeable future via some transitional path (cf. Gheaus, 2013; Gilabert, 2012*b*). For the curious, Wiens (forthcoming*b*) presents my preferred analysis of feasibility.

that best realize Rawls's basic evaluative criteria. Suppose, following most proponents of the target view, that the appropriate target for real world political reform is $B(F)$, the set of morally optimal feasible worlds. The principles we should aim to satisfy are thus the set of directive principles that codifies the normatively constitutive features of the set of optimal feasible worlds. Let P^F denote this set of principles.

Per the model above, let R denote Rawls's assumptions about the empirical constraints on the extent to which our evaluative criteria can be jointly realized. $W(R)$ thus denotes the set of worlds at which Rawls's assumed constraints are satisfied and $B(R)$ denotes the set of optimal worlds given Rawls's assumed empirical constraints. Let P^R denote Rawls's set of directive principles (the two principles of justice and the priority rules). Given section 2, P^R is meant to optimally reflect Rawls's basic evaluative criteria in light of his assumed empirical constraints.

The target view is vindicated if we can show that we can justifiably expect Rawls's principles to characterize the set of optimal feasible worlds ($P^R = P^F$). The target view is defeated if we show that Rawls's principles play no informative role in our reasoning to P^F. Notice that Rawls's assumed constraints accurately specify the feasibility constraints that obtain in the actual world or they do not; that is, $W(R) = W(F)$ or $W(R) \neq W(F)$. I now demonstrate that, assuming either disjunct, Rawls's political ideal plays no informative role in our reasoning to nonideal directive principles — Rawls's ideal qua target is uninformative for the purposes of nonideal theory.

Suppose $W(R) = W(F)$. Then, as a matter of fact, $B(R) = B(F)$ and, thus, $P^R = P^F$. In words: if Rawls's constraints accurately specify the feasibility constraints that obtain at the actual world, then, in fact, Rawls's principles characterize the set of optimal feasible worlds.[21] However, we are unwarranted in simply assuming that Rawls's constraints accurately specify the feasibility constraints that obtain at the actual world, i.e., that $W(R) = W(F)$. This thought is justified only by a reasonably thorough analysis of the relevant facts at the actual world. Such an analysis includes (at least): a diagnosis of the causal mechanisms that generate the outcomes we actually observe and the various ways in which the specified causal mechanisms can be altered given the status quo (see Wiens, 2013). So, without simply assuming that $W(R) = W(F)$, we must justify a reasonable expectation that $B(R) = B(F)$. We can justify this expectation only if we first identify $B(F)$ and compare it with $B(R)$. But then analyzing $B(R)$ in addition to analyzing $B(F)$ is redundant. Analyzing $B(F)$ is sufficient to identify P^F, the directive principles we ought to satisfy. Most importantly, analyzing $B(R)$ (and specifying P^R) in addition to analyzing

21 This thought resonates with Rawls's (2001, p. 4) phrase "realistic utopia", which many philosophers use to characterize the point of analyzing political ideals.

$B(F)$ could not possibly tell us anything about P^F that we don't already glean from analyzing $B(F)$. So, even if $P^R = P^F$ in fact, we can't make justified inferences from P^R to P^F without first identifying $B(F)$ and comparing it with $B(R)$. Thus, if $W(R) = W(F)$, analyzing $B(R)$ is *uninformative* for the purposes of specifying P^F.[22]

This argument is underwritten by the claim that inferences from Rawls's political ideal to nonideal directive principles are unwarranted without first justifying a reasonable expectation that $B(R)$, the set of morally optimal worlds given Rawls's assumed empirical constraints, is sufficiently similar to $B(F)$, the set of morally optimal worlds given the feasibility constraints that obtain at the actual world. Given the need to identify $B(F)$ in justifying this expectation, we have no reason to analyze $B(R)$; analyzing $B(F)$ is enough to identify the directive principles that constitute the appropriate normative target.

Can we nonetheless proceed to analyze political ideals on the assumption that the optimal feasible world is the closest approximation to the political ideal (e.g., Gilabert, 2012*b*, p. 243)? If so, then nonideal directive principles should approximate Rawls's principles as closely as possible and, thus, Rawls's political ideal is informative for nonideal theory. To answer this question, suppose the second disjunct above: $W(R) \neq W(F)$. As above, we have no reason to think, without empirical investigation, that $B(R) = B(F)$, nor that $P^R = P^F$. It follows that analyzing Rawls's political ideal is informative for nonideal theory only if we can justify a reasonable expectation that P^F is the closest feasible approximation to P^R (more accurately: that P^F characterizes the set of worlds in $W(F)$ whose normatively constitutive features are most similar to those at $B(R)$).

First, some more formal machinery. Let's say that a world, w, is consistent with a principle, p, if and only if p is satisfied at w; let $f(p)$ denote the set of worlds that are consistent with p (where f is a function from principles to worlds). Two principles, p and p' are compossible if and only if there is a world, w, that satisfies both p and p'. Let P denote an arbitrary set of principles; $f(P)$ denotes the set of worlds that satisfy all the principles in P. Let $P^R(P) \subseteq P^R$ denote a subset of Rawls's principles that are jointly compossible with P.[23] With all this to hand, let's say that P is *content-wise similar* to Rawls's ideal principles to the extent that P^R is compossible with P. The degree of compossibility with P is measured by the size of the largest subset of Rawls's principles

22 Let me reiterate an important caveat here. I do not propose that nonideal theorists should aim to identify and analyze $B(F)$ and, hence, that P^F specifies the nonideal directive principles we should actually aim to satisfy. In fact, I am pessimistic about the prospects for locating the feasibility frontier and, hence, about the prospects for characterizing the set of morally optimal feasible worlds (see Wiens, forthcoming*b*, secs. 4 and 5). My only claim here is that political ideals are unhelpful for specifying nonideal directive principles, even if we (optimistically) assume that we can locate the feasibility frontier.

23 Precisely: $f(P) = \bigcap f(p_i), p_i \in P$; and $P^R(P) = \{p^R \in P^R : \bigcap_i f(P \cap p_i^R) \neq \varnothing\}$.

that are jointly compossible with P, which we denote (with some abuse of notation) $\max|P^R(P)|$. Accordingly, P is content-wise similar to P^C to a greater degree than P' if and only if $\max|P^R(P)| > \max|P^R(P')|$.

Now take an arbitrary subset of non-optimal feasible worlds, $W_i \subseteq W(F)\backslash B(F)$, and let P^i denote the set of principles that codifies the normatively constitutive features of W_i. We have reason to think that P^F is the closest feasible approximation to Rawls's political ideal — that we can make justifiable inferences from P^R to P^F — only if we can justify a reasonable expectation that there is no P^i that is more content-wise similar to P^R than P^F. We can justify such an expectation only if we have specified P^F and compared its degree of content-wise similarity to P^R with P^i's degree of content-wise similarity to P^R for numerous W_i.[24] So we can justify a reasonable expection that P^F is the closest feasible approximation to P^R only if we specify P^F, which requires identifying and analyzing $B(F)$. Analyzing $B(F)$ is sufficient to identify P^F, the directive principles we should aim to satisfy in our world. Moreover, analyzing $B(R)$ (and specifying P^R) in addition to analyzing $B(F)$ could not possibly tell us anything about P^F that we don't already glean from analyzing $B(F)$. So, if $W(R) \neq W(F)$, analyzing $B(R)$ is *uninformative* for the purposes of specifying P^F.

This argument is underwritten by two claims. The first is that we cannot justifiably expect that we should try to satisfy Rawls's principles as closely as possible without first showing that, among all the sets of principles that characterize subsets of feasible worlds, the set of principles that characterizes the set of optimal feasible worlds, P^F, is most content-wise similar to Rawls's ideal. The second is that we can't show that P^F is most content-wise similar to Rawls's ideal until we have (at least) identified and analyzed the normatively constitutive features of $B(F)$, the set of optimal feasible worlds. But then we have no reason to analyze the normatively constitutive features of $B(R)$, the set of optimal worlds given Rawls's assumed empirical constraints.

In sum, Rawls's assumed constraints accurately specify the feasibility constraints that obtain in the actual world or they do not. In either case, Rawls's political ideal plays no informative role in our reasoning to appropriate nonideal directive principles. More pointedly, political ideals specified without regard for the feasibility constraints that obtain in our world do not present justifiable targets for real world reform. The target view is thus defeated.

24 The point here resonates with that demonstrated by the "general theory of second best" (Lipsey and Lancaster, 1956).

5. AGAINST IDEAL BENCHMARKS

According to the benchmark interpretation of the ideal guidance view, ideal principles like Rawls's do not offer moral directives for our world (even if they are understood as the directive principles that characterize a fully just state of affairs) — not directly anyway. Instead, they provide moral standards by which we comparatively evaluate our feasible options. Our analysis of the feasible options that rank best in light of ideal principles yields nonideal directive principles. So we need an analysis of the political ideal — of the directive principles satisfied at a fully just state of affairs — to help us pick out the feasible options from which to derive nonideal directives.

If political ideals are to serve this evaluative function, then they must plausibly serve to rank possible worlds across the feasible set, not solely among a subset of feasible worlds. In this section, I argue that ideal principles are ill-suited to perform this evaluative function because their range of application is too limited. (As in the last section, I state the argument in terms of Rawls's theory.)

Let's get clear on what's at issue here before we proceed. There are now two distinct concepts to which the adjective "evaluative" might be applied in this section: basic evaluative criteria and (ideal-directive-cum-)evaluative principles. As I define the notion in section 3, a basic evaluative criterion consists of a representation or an analysis of an abstract moral or social value, where basic values are things like freedom, equality, security, peace, community, and so on. A set of basic evaluative criteria identifies the values we use to comparatively evaluate possible worlds and specifies their content. (I've given several examples above to illustrate the proposal.) The question we are considering here is whether principles like Rawls's two principles are best understood as evaluative principles. That is, we want to inquire whether a set of directive principles that characterizes a fully just state of affairs can, following Hamlin and Stemplowska's proposal, serve as evaluative criteria as I define the latter role.

Put this way, the answer might seem to be a resounding "no!" After all, per the model in section 3, Rawls's two principles are not analyses of basic values like freedom or equality; they are attempts to codify the institutional implications of our commitment to freedom and equality, modeled as they are by the original position. In other words, Rawls's two principles are not basic in the way that evaluative criteria are (cf. Cohen, 2003; Swift, 2008).[25] Moreover, given the model, using ideal directive principles to compare and rank possible worlds is redundant — our set of evaluative criteria already does the job.

25 Although I don't follow Cohen here in arguing that non-basic principles are not proper principles of *justice*. We can defuse this dispute between Cohen and Rawls by simply noting that Cohen is ostensibly concerned with evaluative principles while Rawls is ostensibly concerned with directive principles (cf.

This might seem too easy. Instead of rejecting the claim that Rawls's principles are apt to serve as evaluative criteria, we might remain convinced that they are fit for the job and take the preceding point as a reason to reject the model in section 3 instead. So let's look for an independent reason (independent of the model, that is) for denying the thought that Rawls's principles can serve a basic evaluative function.

Start by considering the evaluative principle implied by Rawls's difference principle: all else equal, w is preferred (from the standpoint of morality) to w' if socioeconomic inequalities are arranged so that the prospects of society's least advantaged members are greater at w than at w', for all $w, w' \in W(F)$ (cf. Sangiovanni 2008, p. 224; Valentini 2011, p. 308).[26] Let x_1 denote the distributive share of society's most advantaged members and x_2 denote the distributive share of society's least advantaged members.[27] Let $W^D \subset W(F)$ denote a subset of feasible worlds such that, for all worlds in W^D, x_2 is maximized when x_1 is somewhere between $20x_2$ and $30x_2$.[28] Let's concede that, for all $w, w' \in W^D$, w is preferred to w' if the prospects of the least advantaged are greater at w than at w', all else equal. This amounts to agreeing that the difference principle yields a plausible ranking for worlds in W^D.

Now let's relax Rawls's assumptions of a closed society. Let $W^O \subset W(F)$ denote the subset of feasible worlds at which: capital can move across international borders relatively freely; it is quite costly for labor to move across borders; and the economic elite (credibly) threaten to transfer their investments abroad unless their tax rates are lowered and the relevant regulations are altered to maximize their profit (so we're also relaxing the assumption that all citizens have a "sense of justice"). Hence, elites use their relative mobility as bargaining leverage to extract a much greater share of total social production, so that x_2 is now maximized when x_1 is somewhere between $125x_2$ and $150x_2$ for all worlds in W^O.[29] So, by assuming capital mobility, relative labor immobility, and profit-maximizing elites, we drastically increase both socioeconomic inequality and the

Mason, 2012). So long as we're clear about this distinction, there's no reason to reserve the term "justice" for one or the other (cf. Wiens, forthcoming*a*).

26 "All else equal" is meant to accommodate Rawls's priority rules by ensuring that worlds are compared with respect to the difference principle only if they share sufficiently similar systems of basic liberties and opportunities.

27 It might be helpful here to recall the figures Rawls uses to illustrate the implications of the difference principle (Rawls, 1999, p. 66).

28 By comparison, the executive-to-worker pay ratio in Norway in 2011 was 58:1; in Japan, it was 67:1 and in Sweden, it was 89:1 <http://www.aflcio.org/Corporate-Watch/Paywatch-Archive/CEO-Pay-and-You/CEO-to-Worker-Pay-Gap-in-the-United-States/Pay-Gaps-in-the-World>.

29 By comparison, the executive-to-worker pay ratio in the United States in 2013 was 331:1 <http://www.aflcio.org/Corporate-Watch/Paywatch-2014>.

	Principle	Least advantaged	Most advantaged
w_1	Difference principle	x_2	$130x_2$
w_2	Stricter equality	$\frac{9}{10}x_2 (= y)$	$45x_2 (= 50y)$

Table 1. Distributive shares under two different principles

prospects of the most advantaged while the prospects of the least advantaged remain fixed.

Surely there are worlds in the feasible set (if not the actual world) at which capital is relatively mobile, economic elites are profit maximizers, and labor is relatively immobile. If the difference principle is to serve a basic evaluative function, as the benchmark view claims, it must buttress a plausible ranking of worlds in W^O too, not only among the worlds in W^D. Does it?

To investigate this question, consider two worlds in W^O, denoted w_1 and w_2. Suppose the difference principle is satisfied (i.e., x_2 is maximized) at w_1 while much stricter limits are placed on inequality at w_2, such that x_2 is not maximized at w_2. The relative distributive shares for the two worlds are as summarized in table 1 (with all shares indexed to the share of the least advantaged in w_1).[30] (Suppose that w_1 and w_2 are equal in all other respects, *mutatis mutandis*; in particular, they both satisfy the equal basic liberties principle and fair equality of opportunity principle.) According to table 1, the share of the least advantaged is slightly less valuable at w_2 than at w_1. Thus, if we evaluate these worlds according to the difference principle, w_1 is (morally) preferred to w_2.

This verdict seems deeply counterintuitive given the basic values Rawls's principles are meant to reflect. In view of w_2, an institutional scheme that satisfies the difference principle (as in w_1) publicly surrenders to the threats of a class of mercenary elites in exchange for a modicum of gain for the least advantaged. In so doing, such an institutional scheme acquiesces in the creation of an economic hierarchy that seems in tension with Rawls's commitment to the moral equality of citizens and a stable system of cooperation for mutual advantage. By contrast, w_2 represents a society that publicly refuses to allow an advantaged elite to dictate the terms of economic cooperation to the rest and, thereby, undermines economic hierarchies and better expresses respect for all citizens' freedom and equality. Accordingly, I submit that w_2 is reasonably preferred to w_1 in light of Rawls's

30 Imagine that Rawls's "contribution curve" for these worlds is fairly flat (Rawls, 1999, p. 66).

basic evaluative criteria; that is, a principle of stricter equality is reasonably preferred to the difference principle given capital mobility and acquisitive elites. The difference principle qua evaluative criterion can't deliver this judgment. Importantly, the foregoing comparative evaluation need not draw on latent worries that the inequality in w_1 will negatively affect citizens' basic liberties or their opportunity sets. The situation described in w_1 is fully compatible with the least advantaged citizens retaining the full fair value of their basic liberties and having fair access to positions within the elite class (I elaborate this point in a moment).

Let's concede that the difference principle yields plausible evaluative judgments for some subsets of feasible worlds (e.g., W^D). Yet a proper set of evaluative criteria must help us rank worlds *across the entire feasible set*, $W(F)$. Since there is a subset of feasible worlds for which the difference principle implies an apparently counterintuitive ranking in view of the values the difference principle purports to reflect — $\{w_1, w_2\}$ — the difference principle is not apt to help us rank our options across the feasible set, as the benchmark view proclaims.

One might think that the difference principle can at least tell us that worlds in W^D are morally superior to worlds in W^O. But this isn't necessarily so. Take two worlds, $w \in W^D$ and $w' \in W^O$, such that the prospects of the least advantaged in both w and w' are maximized at the same level, x_2. Suppose that individuals across both worlds enjoy the same system of basic liberties and the same level of equality of opportunity. There is apparently little reason to think that capital mobility necessarily undermines equal basic liberties or equality of opportunity. To ensure this, assume the following. First, the differences in inequality across w and w' (say, 30:1 versus 125:1) are not due to differences in the distribution of political influence, but differences in investment decisions — inequality is greater at w' than at w because cross-border economic interactions increase social efficiency but economic elites use their leverage to capture all the efficiency gains. Second, all individuals are educated in public schools of uniform quality and private education is prohibited, so that the least advantaged have effectively equal opportunities to join the economic elite. Given these assumptions, the difference principle qua evaluative criterion does not rank $w \in W^D$ above $w' \in W^O$ even though inequality is much higher at w'. (As I've stated the evaluative criterion implied by the difference principle, greater prospects for the least advantaged at w than at w' is only a sufficient condition for ranking w above w'. If greater prospects for the least advantaged is also a necessary condition, then the difference principle implies — counterintuitively, I submit — that we should be indifferent between $w \in W^D$ and $w' \in W^O$.)[31]

31 This result depends on how we interpret the way in which the difference principle handles pairs of cases

Although the preceding argument does not depend on the model developed in section 3, that model offers a plausible explanation for why Rawls's principles are ill-suited to help us rank options across the entire feasible set. Ideal principles are "point solutions" — they codify the practical implications of our commitment to certain basic values *at a particular point within a (constrained) set of possibilities*. As Rawls puts it, "the choice of the principles of justice presuppose[s] a certain theory of social institutions" (Rawls, 1999, p. 138); that is, the parties in the original position rank alternative sets of principles according to how well their satisfaction would realize the basic evaluative criteria modeled by the original position *in view of certain facts about how the world operates*. Indeed, it can't be otherwise — the parties have no basis for choosing one set of principles over another unless they understand how the benefits and burdens of social cooperation will be distributed by institutional schemes that satisfy alternative sets of principles. This requires understanding how individuals are likely to respond to the different norms and incentives generated by alternate institutional schemes; for example, how different kinds of institutions shape individuals' labor, consumption, investment, and savings behavior. But, as the discussion of table 1 shows, there is no guarantee that our basic values, which are best reflected by the difference principle given one set of facts, won't be better reflected by a principle requiring stricter socioeconomic equality given a different set of facts. I put the point more generally in section 4: we cannot justify a reasonable expectation that the normative principles that best reflect our basic values given one set of facts will, given a different set of facts, reflect our basic values to a greater degree than a content-wise dissimilar set of principles. Thus, as point solutions, political ideals — the directive principles that characterize a fully just state of affairs — are ill-suited to serve the evaluative purpose of ranking our options across the full range of feasible worlds. At best, they are apt to mislead in this regard. The benchmark view is thus defeated.

6. CONCLUSION

In section 3, I develop a model of normative political theorizing according to which political ideals codify the normatively constitutive features of the worlds that best realize

of the following sort: the least advantaged have the same prospects in each (say, x_2) but the most advantaged get differing amounts (say, $50x_2$ and $75x_2$). As Van Parijs (2003) shows, Rawls's arguments don't decide the matter. As I understand Van Parijs's treatment of the issue, the way to determine the implications of the difference principle in the type of case I've given in the text is to enrich the set of basic criteria by which we evaluate the possibilities. This is consistent with my claim that we evaluate the possibilities according to evaluative criteria that are more basic than the difference principle. (Thanks to an anonymous reviewer for raising the issue.)

our basic values given a set of empirical constraints on the joint realization of our basic values. This model is a generalization of insights gleaned (in section 2) from the structure of Rawls's and Nozick's arguments for their favored principles. I then use this model to present arguments against the two interpretations of the ideal guidance view: the target view (section 4) and the benchmark view (section 5). Given these arguments, I conclude that the prevailing ideal guidance view is defeated — political ideals contribute nothing to (and can even mislead) our reasoning to a set of directive principles that are suited for unjust or otherwise nonideal circumstances.

I reiterate two important caveats. First, my argument is not about how best to do nonideal theory. Indeed, elsewhere I have expressed pessimism about our ability to estimate the feasibility frontier (Wiens, forthcoming*b*). Instead, my argument implies that, however we should do nonideal theory, it should not be with an eye to political ideals. Second, my argument does not imply that ideal theory should be wholly abandoned (as other skeptics apparently suggest); for example, exploration of political ideals might help motivate agents to pursue feasible reforms that can lead to improvements from the standpoint of justice (cf. Gheaus, 2013). So we may yet have reasons to investigate political ideals. Investigating potentially fruitful uses of political ideals is an area where much research remains to be done.

I conclude with a brief comment regarding the relationship between normative political philosophy and social science. According to the model developed in section 3, our reasoning to a set of directive political principles for our nonideal world has two equally basic inputs: a set of moral evaluative criteria and a specification of the feasible set. The fundamental importance of moral considerations assures abstract political philosophy of an important role in normative theorizing (Swift, 2008). But the fact that feasibility considerations are on a par with moral considerations in the model shows that social scientists play an equally important role in normative political theory. Contra the traditional division of labor, social science does not enter the picture *after* we have specified directive principles, to help with implementation (cf. Miller, 2008). Instead, social science enters alongside moral theory to help with the specification of directive principles suited for nonideal circumstances. The reasoning here is simple. We cannot identify the appropriate target for real world reform (i.e., the directive principles we should aim to eventually satisfy) until we can identify the optimal world relative to an appropriate set of empirical constraints.[32] We cannot identify the optimal world relative to the specified constraints until we identify the bounds of the constraint set. We cannot identify the

32 Here I leave open which sets of constraints qualify as "appropriate" for normative political theorizing; but see (Wiens, forthcoming*b*, sec. 5).

bounds of the constraint set without rigorous social science. Thus, we cannot identify the appropriate normative target for political reform without rigorous social science — social scientists have an integral role to play in judicious normative political theory. This means that political philosophers can ill-afford to neglect collaboration with social scientists in specifying directive political principles. But it also means that there is a great need for social scientists to expand the scope of their attention, from explanation of the status quo to locating the limits of political possibility. Social scientists have an essential contribution to make to the normative theoretical enterprise.

REFERENCES

Charlow, Nate. forthcoming. Decision Theory: Yes! Truth Conditions: No! In *Deontic Modals*, ed. Nate Charlow and Matthew Chrisman. Oxford: Oxford University Press.

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

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

Geuss, Raymond. 2008. *Philosophy and Real Politics*. Princeton and Oxford: Princeton University Press.

Gheaus, Anca. 2013. "The Feasibility Constraint on the Concept of Justice." *The Philosophical Quarterly* 63(252):445–464.

Gilabert, Pablo. 2012*a*. "Comparative Assessments of Justice, Political Feasibility, and Ideal Theory." *Ethical Theory & Moral Practice* 15(1):39–56.

Gilabert, Pablo. 2012*b*. *From Global Poverty to Global Equality*. New York: Oxford University Press.

Hamlin, Alan and Zofia Stemplowska. 2012. "Theory, Ideal Theory and the Theory of Ideals." *Political Studies Review* 10:48–62.

Jubb, Robert. 2012. "Tragedies of Non-Ideal Theory." *European Journal of Political Theory* 11(3):229–246.

Kratzer, Angelika. 1991. Modality. In *Semantics: An International Handbook of Contemporary Research*, ed. Arnim von Stechow and Dieter Wunderlich. Berlin: de Gruyter pp. 639–650.

Lewis, David. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lipsey, R.G. and Kelvin Lancaster. 1956. "The General Theory of Second Best." *The Review of Economic Studies* 24(1):11–32.

Mason, Andrew. 2012. "What is the Point of Justice?" *Utilitas* 24(4):525–547.

Miller, David. 2008. Political Philosophy for Earthlings. In *Political Theory: Methods and Approaches*, ed. David Leopold and Marc Stears. New York: Oxford University Press.

Mills, Charles W. 2005. "'Ideal Theory' as Ideology." *Hypatia* 20(3):165–184.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Pogge, Thomas W. 1989. *Realizing Rawls*. Ithaca, NY: Cornell University Press.

Rawls, John. 1999. *A Theory of Justice*. 2 ed. Cambridge, MA: Harvard University Press.

Rawls, John. 2001. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.

Robeyns, Ingrid. 2008. "Ideal Theory in Theory and Practice." *Social Theory and Practice* 34(3):341–362.

Sangiovanni, Andrea. 2008. Normative Political Theory: A Flight From Reality? In *Political Thought and International Relations: Variations on a Realist Theme*, ed. Duncan Bell. Oxford: Oxford University Press pp. 219–239.

Schmidtz, David. 2011. "Nonideal Theory: What It Is and What It Needs to Be." *Ethics* 121(4):772–796.

Sen, Amartya. 2009. *The Idea of Justice*. Cambridge, MA: Harvard University Press.

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

Stemplowska, Zofia. 2008. "What's Ideal About Ideal Theory?" *Social Theory and Practice* 34(3):319–340.

Swift, Adam. 2008. "The Value of Philosophy in Nonideal Circumstances." *Social Theory and Practice* 34(3):363–387.

Valentini, Laura. 2009. "On the Apparent Paradox of Ideal Theory." *Journal of Political Philosophy* 17(3):332–355.

Valentini, Laura. 2011. "A Paradigm Shift in Theorizing About Justice? A Critique of Sen." *Economics and Philosophy* 27:297–315.

Valentini, Laura. 2012. "Ideal vs. Non-ideal Theory: A Conceptual Map." *Philosophy Compass* 7(9):654–664.

Van Parijs, Philippe. 2003. Difference Principles. In *The Cambridge Companion to Rawls*, ed. Samuel Freeman. Cambridge: Cambridge University Press.

Wiens, David. 2012. "Prescribing Institutions Without Ideal Theory." *The Journal of Political Philosophy* 20(1):45–70.

Wiens, David. 2013. "Demands of Justice, Feasible Alternatives, and the Need for Causal Analysis." *Ethical Theory & Moral Practice* 16(2):325–338.

Wiens, David. forthcoming*a*. "'Going Evaluative' to Save Justice From Feasibility — A Pyrrhic Victory." *The Philosophical Quarterly* .

Wiens, David. forthcoming*b*. "Political Ideals and the Feasibility Frontier." *Economics and Philosophy* .

Williams, Bernard. 2005. *In The Beginning Was The Deed: Realism and Moralism in Political Argument*. Princeton: Princeton University Press.