Version 5.5 // Nov 2015

Assessing Ideal Theories: Lessons from the Theory of Second Best

David Wiens

1. INTRODUCTION

Normative political theorizing in the wake of Rawls (1971/1999) has been preoccupied with analyzing with analyzing the constitutive features of ideally just states of affairs ("political ideals").[1] These so-called "ideal theories" typically assume idealized social conditions in which people are to coordinate social activity and resolve social conflicts; for example, societies are assumed to be closed to cross-border interactions (Rawls, 1999), or certain kinds of market failures are assumed away (Nozick, 1974). There is growing debate about the virtues and vices of such ideal theorizing, with much of this debate focusing specifically on the merits of normative theories that assume idealized or "unrealistic" background social conditions (see, among others, Farrelly, 2007; Mills, 2005; Stemplowska, 2008; Valentini, 2009). In the context of this debate, several philosophers have thought that conventional ideal theories face a challenge from the "general theory of second best" (e.g., Brennan and Pettit, 2005; Estlund, 2008; Goodin, 1995; Mason, 2004; Swift, 2008). As Lipsey and Lancaster (1956, hereafter L&L) originally put it, "The general theorem of the second best states that if one of the Paretian optimum conditions cannot be fulfilled a second best optimum situation is achieved only by departing from all other optimum conditions" (p. 12).[2] The predominant transposition by political philosophers says (roughly): if one of the background social conditions assumed when analyzing a

Author's note. Earlier versions of this paper were presented at the Australian National University and at the MANCEPT Workshops at the University of Manchester. Thanks to those audiences, and in particular to Marcus Arvan, Geoff Brennan, Dave Estlund, Bob Goodin, Holly Lawford-Smith, and Nic Southwood for helpful discussion. Thanks also to the Editor, Andrew Williams, and anonymous referees for their comments. Research for this paper was supported by ARC Discovery Grant DP120101507.

1 Hereafter, I use "political ideal" ("ideal" for short) to denote a set of normative principles that specifies certain constitutive features of a fully just state of affairs. To "realize an ideal" is to realize a state of affairs that fulfills these principles. An "ideal theory" presents an analysis of a political ideal.

2 See Ng (2004, ch. 9) for an accessible introduction. I'll use "theorem" to refer to the centerpiece of L&L's original paper, viz., the theorem proved in section 7 of that paper; I'll use "theory" to refer to the entity composed of the theorem, its proof, and the implications drawn therefrom.

Penultimate version. To appear in *Politics, Philosophy, and Economics*.

political ideal does not obtain, then the (normatively) best state of affairs under the circumstances does not necessarily satisfy the principles that characterize a fully just state of affairs. Since theoretical analyses of political ideals typically assume background conditions that seem infeasible or unlikely to obtain at the actual world (as noted above), we should not necessarily aim to satisfy the principles that characterize a fully just state of affairs as far as possible (assuming we should aim to realize the normatively best state of affairs under any given circumstances). Thus understood, the theory of second best subverts one of the primary rationales for philosophers' preoccupation with analyzing political ideals — namely, to specify (long-range) targets for real world reform efforts (see, e.g., Rawls, 1999; Robeyns, 2008; Simmons, 2010; Valentini, 2009).[3]

In this paper, I show that philosophers have generally misinterpreted the second best theorem, its antecedent in particular. Thus, they have misunderstood the nature of the challenge it raises. L&L's theorem applies, not when one of the idealized conditions typically assumed by ideal theories fails to obtain, but *when one of the principles that characterize an ideal state of affairs fails to be satisfied*. Transposed more accurately, the second best theorem says (roughly): if one of the principles that characterize a fully just state of affairs remains unsatisfied, then the best state of affairs under the circumstances does not necessarily satisfy as many of the remaining principles as possible. Consequently, the theory of second best presents a double-edged sword in the context of recent methodological debates.

On the one hand, the theory of second best challenges the prevailing view that political ideals provide appropriate targets for real world reform efforts *only if we have good reason to expect that ideal normative principles will remain unsatisfied.* Given that the italicized condition is highly controversial, the theory of second best hardly provides a compelling challenge to the general view that political ideals can serve as normative targets for real world reforms. On the other hand, the theory of second best poses a stiff *anti-approximation warning*: if a political ideal remains unrealized, then an approximate realization of that ideal is not necessarily the best thing under the circumstances. Put differently, we cannot expect steady progress in the direction of the ideal to yield improvements from the standpoint of justice (*pace*, e.g., Christiano and Braynen 2008; Gilabert 2012*b*, p. 243; Valentini 2012*b*, p. 42). Accordingly, as I will show, the theory of second best requires ideal theorists to undertake certain kinds of causal and comparative analyses

3 Alternative rationales for doing ideal theory are on offer; for example, that analyzing the constitutive features of ideally just states of affairs yields normative criteria for evaluating and selecting feasible institutional schemes (see, e.g., Gilabert, 2012*a*; Swift, 2008). For critical discussion of this "benchmark view" of ideal theory, see Wiens (forthcoming*a*).

that are thought to lie beyond the remit of conventional ideal theory. Thus, while the theory of second best fails to challenge the general claims that ideal theories can present appropriate reform targets, it does not leave the conventional practice of specifying these normative targets to proceed untouched.

My argument begins in section 2 with a brief introduction to the theory of second best; section 3 continues by surveying extant applications of the theory of second best to normative political theory and outlines what's at stake for political philosophers in interpreting the theorem correctly. Section 4 discusses the theorem in more detail, while section 5 discusses its implications for ongoing methodological debates.

Let me register a caveat before I continue. Although welfare economics provides the rhetorical context in which L&L prove their theorem, their result is general enough to apply to "all maximization problems, not just welfare theory" (L&L, 12, n. 2). One might deny that the theory of second best has any implications for normative political theory by arguing that the latter is not the relevant sort of maximization exercise.[4] Whether normative political analysis is isomorphic to the sort of maximization exercise presupposed by the theorem's proof is an issue that has been little discussed among philosophers. Alas, this issue must remain beyond the scope of this paper. I will follow extant philosophical discussion of the second best in assuming that the theorem applies to normative political theory.[5] Given this assumption, my aim is to present an accurate interpretation of the conditions for its application and thereby illuminate its implications for ongoing philosophical debates.

2. THE THEORY OF SECOND BEST: AN INTRODUCTION

At face value, L&L's theorem bears on debates in welfare economics. According to one stereotypical view from economics, an ideal economic system is one that allocates economic goods *Pareto efficiently*. An allocation of economic goods is Pareto efficient if and only if there are no transactions to be made that would increase at least one individual's

4 Although this would not to show that the anti-approximation point implied by the theorem fails to apply to normative political theory. It would only show that L&L's proof, which presupposes a particular sort of analytic framework, does not support the application to political theory. The anti-approximation point might be established via other avenues (see, e.g., Wiens, forthcoming*a*).

5 Several philosophers have suggested the required isomorphism; among others, see: Cohen (2008); Goodin (1995); Hamlin and Stemplowska (2012). Wiens (forthcoming*a*) discusses the issue at length, using Rawls's and Nozick's theories of justice to show that familiar normative theories are fruitfully modeled as the relevant sort of maximization exercises.

utility without also decreasing another's utility.[6] Using the techniques of mathematical optimization, we can derive general constitutive conditions for a Pareto efficient allocation.[7] First, for all pairs of goods consumed in the economy, the rate at which individuals are willing to trade one good for another (the marginal rate of substitution in exchange, or MRS) must be the same for all individuals who consume those goods; second, for all pairs of goods produced in the economy, the rate at which one input can be substituted for another in production (the marginal rate of technical substitution, or MRTS) must be the same for all producers who use those inputs; third, for any pair of goods, the MRS for that pair must equal the rate at which the economy can redirect production of one good into production of the other (the marginal rate of transformation, or MRT). These three conditions specify the "optimum conditions" for an ideal economic system; they are constitutive conditions of the efficient ideal.

Let's make this a little less abstract. Suppose we have an economy with three goods: apples, bananas, and coconuts. Let $p_1 = \$3$ be the price of a pound of apples and $p_2 = \$1$ be the price of a pound of bananas (set aside coconuts for now). Assuming all consumers face the same prices, $p_1/p_2 = 3$ is the MRS for apples and bananas — each person consumes apples and bananas to the point where he or she is willing to exchange one pound of apples for three pounds of bananas.[8] The MRT is equal to the relative marginal costs of producing each good — the cost of producing one more pound of apples (MC_1) over the cost of producing one more pound of bananas (MC_2). If the allocation is Pareto efficient, then prices are equal to the producers' marginal cost: $p_1 = MC_1 = \$3$ and $p_2 = MC_2 = \$1$. So the MRT is equal to $MC_1/MC_2 = 3$. Thus, if the allocation is Pareto efficent, consumers' MRS equals the MRT. (If p_i and p_j denote prices for production inputs — labor and fertilizer, say — then producers' MRTS must also be $p_i/p_j = 3$ if the allocation is Pareto

6 More precisely: If x denotes the status quo allocation of goods, $u_i(x)$ denotes the utility i receives from x, and N denotes the number of individuals in the economy, then x is Pareto efficient if and only if there is no alternative allocation y that satisfies both of the following conditions.

(1) $u_i(y) \geq u_i(x)$ for all $i = 1, 2, \ldots, N$

(2) $u_i(y) > u_i(x)$ for at least one i.

7 See Ng (2004, ch. 2) for an accessible derivation of these conditions.

8 But people don't all consume the same quantities of apples and bananas, to be sure; in particular, they don't necessarily consume apples and bananas at a rate of 3 to 1. To wit, Josh might consume a basket consisting of 5 pounds of apples and 1 pound of bananas, while Donna consumes a basket consisting of 1 pound of apples and 2 pounds of bananas. Yet, in an efficient equilibrium, they have the same MRS — both Josh and Donna are willing to trade 3 pounds of apples for 1 pound of bananas. Thus, Donna is indifferent between, on the one hand, a basket with 1 pound of apples and 2 pounds of bananas and, on the other hand, a basket with 4 pounds of apples and 1 pound of bananas.

efficient.)

Since we live in a world that typically violates these optimum conditions, a pressing question arises: what should we do if one or more of these conditions will remain unsatisfied? Suppose, for instance, that there is a coconut monopoly, which restricts the supply of coconuts to drive up the price. Hence, the price of coconuts is higher than the marginal cost of producing coconuts, $p_3 = \$3 > MC_3 = \$1.50$. Since $p_1/p_3 = 1$, people will consume apples and coconuts at the point where they are willing to exchange one coconut for one pound of apples. But, notice that $MC_1/MC_3 = 2$, so consumers' MRS for apples and coconuts is not equal to producers' MRT for those goods — allocative efficiency fails to obtain. To wit, notice that $p_1/p_3 = 1$ implies that consumers are indifferent between one pound of apples and one coconut. But $MC_1/MC_3 = 2$ implies that the economy can produce two coconuts at the same cost as producing one pound of apples. Together, these ratios imply that we can increase everyone's welfare by shifting some productive capacity from coconuts to apples.[9] (Analogous results hold if we substitute bananas for apples.)

If the coconut monopoly remains in place and $p_3 \neq MC_3$, then, in our three good economy, the MRS and MRT can be equal for at most one pair of goods, namely, apples and bananas. What should we do if $p_3 \neq MC_3$? One intuitive answer is that we should nonetheless aim to satisfy the conditions for Pareto efficiency as far as possible — that is, we should do what we can to ensure that the MRS for apples and bananas equals the MRT for apples and bananas. The theory of second best is meant to show that this "piecemeal" approach to economic policy is, in general, misguided. Specifically, L&L prove that if the satisfaction of one of the optimum conditions is constrained, then it is not necessarily true that the "second best optimum"[10] — i.e., the best outcome among the remaining possibilities — satisfies the optimum conditions as far as possible. L&L's theorem has an important negative corollary: given a situation where at least one of the optimum conditions remains unsatisfied, we cannot determine a priori whether movement toward greater satisfaction of the optimum conditions constitutes an improvement over the status quo (L&L, 1956, 12). In terms of our simple economy, the theory of second best

9 If no one wants to have two coconuts, they could trade one of them for some apples. So two coconuts is preferable to one coconut even if no one actually wants two coconuts.

10 The phrase "second best" is potentially misleading here. Suppose we have an ordinal ranking of all possible outcomes, with A being the ideal, B being the next best outcome, C being the best after B, and so on. The conditions for a "second best optimum" do not necessarily characterize B, i.e., the next best outcome after the ideal outcome. In general, they characterize the best outcome attainable given that we cannot satisfy the optimum conditions fully. This is consistent with that outcome being ranked (e.g.) 47th in a complete ordinal ranking. "Second best" simply refers to any situation where failure to realize the first best or ideal optimum is taken as a constraint. (Thanks to Dave Estlund for bringing this potential confusion to my attention.)

shows the following. For a world in which the MRS for coconuts and apples fails to equal the MRT for those goods, (1) equalizing the MRS and MRT for apples and bananas is not necessarily the best outcome; and (2) we cannot determine a priori whether equalizing the MRS and MRT for apples and bananas yields an improvement over a status quo in which the MRS and MRT are unequal for all pairs of goods.

3. THE THEORY OF SECOND BEST IN NORMATIVE POLITICAL THEORY

Political philosophers have drawn on the theory of second best to challenge what we might call the "Target View" of ideal theory, which claims that our efforts to organize our collective affairs in the actual world should place us on a transitional path toward eventual realization of a political ideal. In other words, the Target View asserts that our political reform efforts should lead us to establish arrangements that satisfy, as far as possible, the normative principles that characterize constitutive features of fully just states of affairs (see, among others, Rawls, 1999; Robeyns, 2008; Simmons, 2010; Valentini, 2009).[11] For instance, if we take Rawls's two principles to characterize certain constitutive features of fully just states of affairs, then the Target View asserts that we should aim to realize states of affairs that satisfy Rawls's two principles as far as possible.

Philosophers have interpreted L&L's second best theorem in one of three ways to challenge the Target View. One argument deploys a *moral values* interpretation, which says that the theorem applies when we cannot realize some of our basic moral and social values, such as liberty, equality, or security: "*When our ideals* [i.e., values[12]] *cannot all be realized simultaneously*, the general theory of the second best…warns us against assuming naively that it is better to implement more of our [values] rather than fewer (or indeed to implement each of them to a greater rather than lesser degree)" (Goodin, 1995, 54, emphasis added). Perhaps the levels of liberty, equality, fraternity, and material prosperity we enjoy in an ideal world exceed some (relatively high) threshold value (to take Goodin's example). Yet perhaps our actual circumstances are such that we cannot realize all of these values simultaneously (at least not at the level enjoyed at the ideal world). If this is so, it follows from the moral values interpretation that we should not necessarily seek to realize as many of our values as we can or realize them to the greatest extent possible.

11 There are, of course, other purposes an ideal theory might serve (see footnote 4). As I'm not interested in these here, I set aside further discussion of the matter.

12 Goodin's use of "ideal" clearly refers to what we might naturally call basic moral or social values, like liberty, equality, community, and so on. Since I reserve the term "ideal" for a particular technical purpose (see footnote 1), I replace Goodin's use of "ideal" with "value" to avoid confusion.

A second argument calls on an *ideal institutions* interpretation, which says that the theorem applies when we are unable to (fully) implement an institutional scheme designed for ideal conditions: "what the theory of the second best suggests is that... [s]ince *departure from any one condition in the institution used for the model* [of the ideal scheme] means that all the other conditions may not be desirable, it is not clear whether the optimum choice is to get as close to the original as possible, or to construct a completely different institution" (Coram, 1996, 93, emphasis added).[13] Given that an ideal institutional scheme will likely remain unrealized, this interpretation implies that we should not necessarily try to implement an ideal institutional scheme in our nonideal world.

If we interpret the theorem of second best in either of the above ways, then it poses little threat to the Target View. To the moral values interpretation, a proponent of the Target View concedes that we cannot simultaneously realize our basic values at a high level but replies that ideal theory is motivated in part by this recognition. Ideal theory delivers principles, like Rawls's two principles, that help us determine the relative weight we should give to disparate values, thereby enabling us to specify the balance of values we should aim to realize — that is, the ideal balance of values (Gilabert, 2012*a*; Swift, 2008). To the ideal institutions interpretation, a proponent of the Target View replies that we should not try to implement ideal *institutions* in nonideal circumstances; rather, we should try to implement the feasible institutional scheme that best satisfies the *normative principles* that characterize central features of the ideal scheme. There is no inconsistency here because normative principles do not have any particular institutional implications (Valentini, 2011).

The predominant interpretation of the theorem's antecedent avoids these replies. This *background assumptions* interpretation says that the theorem applies when the background social conditions assumed in specifying ideal normative principles do not obtain: "*If any one of the conditions presupposed by ideal theory is missing*, then the Theory of Second-Best warns that we might... need to make systematic alterations right across the board in the prescriptions of ideal theory" (Goodin 2012, 162, emphasis added; cf. Brennan and Pettit 2005; Räikkä 2000). Typically, ideal theory assumes social conditions that are

13 Cf. Brennan and Pettit (2005); Wiens (2012). Coram goes on to claim that the theory of second best warns us of two fallacies: "the fallacy of continuity", which holds that similar initial conditions produce similar results; and "the fallacy of stretchability", which holds that small changes to institutions leads to small changes in the outcome (Coram, 1996, 94). These are surely fallacies, as Coram's examples show; but they apply to the ideal world as much as they apply to the nonideal world. To wit, assuming we can fulfill the conditions required to achieve Pareto efficiency, small changes to the initial economic endowment can produce very different Pareto efficient outcomes. Hence, the insight of the theory of second best is not, *pace* Coram, that "radical alterations in institutions may be required to accommodate small shifts in initial conditions" (Coram, 1996, 91). That insight bears on the ideal case too.

unlikely to obtain in the actual world, abstracting away from factors that constrain our realization of basic moral and social values. For example, when specifying his principles of justice, Rawls notably assumes that individuals have a "sense of justice" that will lead them to refrain from taking advantage of others; that society has sufficient material resources to protect each citizen's basic liberties equally; and that society is self-contained (i.e., there are no cross-border transactions) (Rawls, 1999, 7, 8, 498). Since these circumstances are unlikely to obtain in the actual world, the background assumptions interpretation implies that we should not necessarily aim to satisfy Rawls's principles of justice.

The background assumptions interpretation poses a greater *prima facie* threat to the Target View than the first two. However one defines "ideal theory", it is true that the specification of ideal normative principles characteristically abstracts from certain nonideal features of the actual world, (implicitly) assuming circumstances that are unlikely to obtain (cf. Hamlin and Stemplowska, 2012; Valentini, 2012*a*). Given this, the background assumptions interpretation implies that we should not necessarily aim to satisfy ideal normative principles in a nonideal world. Is this an accurate interpretation of the theorem?

4. INTERPRETING THE THEOREM

Recall L&L's informal statement of the second best theorem: "if one of the Paretian optimum conditions cannot be fulfilled a second best optimum situation is achieved only by departing from all other optimum conditions" (L&L, p. 12).[14] What's at issue above is the interpretation of the antecedent, in particular, the phrase "optimum conditions". At a glance, "optimum conditions" might refer to one of two things in the welfare economics context in which the theorem was introduced. First, the relevant "optimum conditions" might be the constitutive conditions for Pareto efficiency stated above: the identity statements about marginal rates of substitution and transformation and so on. In this case, "optimum conditions" refers to the *output* of a theoretical exercise; they are the result of analyzing an abstract (mathematical) model of the economy. Second, the relevant

14 It is well known among economists that L&L's informal statement is too strong (given what they actually prove), in at least two respects: (1) the theorem holds whether one of the optimum conditions "cannot be fulfilled" or merely remains unfulfilled despite its fulfillment being possible (see footnote 16). (2) "[D]eparting from all other optimum conditions" is not the "only" way to achieve a second best optimum; there are "separability" conditions under which fulfilling the remaining optimum conditions achieves a second best optimum (see, e.g., Blackorby, Davidson and Schworm, 1991; Davis and Whinston, 1965). Put more carefully, L&L should have said "a second best optimum situation is not necessarily achieved by fulfilling all other optimum conditions". Since these interpretive issues are beside my purpose in this paper, I set them aside. I acknowledge them solely to mitigate distraction from the main issue.

"optimum conditions" might be the background conditions assumed by the model used to derive these output conditions: for instance, the absence of production and consumption externalities; that agents have symmetric and perfect information; the absence of monopolies; and costless transactions. In this case, "optimum conditions" refers to the *input* of a theoretical exercise; they specify core features of the economic model to be analyzed.

To avoid ambiguity, let $O = \{o_1, \ldots, o_n\}$ denote an arbitrary set of "optimum conditions" understood as the output of a theory and let $I = \{i_1, \ldots, i_n\}$ denote an arbitrary set of "optimum conditions" understood as the input of a theory. In the welfare economics example given above, O denotes the set of identity statements about marginal rates of substitution and transformation, while I denotes the assumptions of the model used to derive O. Given this distinction (and assuming L&L did not equivocate), we have two possible specifications of the second best theorem:

> *The Theorem of Second Best — Output.* "[I]f one of $[o_1, \ldots, o_n]$ cannot be fulfilled a second best optimum situation is achieved only by departing from all the other $[o_j$ in $O]$."[15]

> *The Theorem of Second Best — Input.* "[I]f one of $[i_1, \ldots, i_n]$ cannot be fulfilled a second best optimum situation is achieved only by departing from all the other $[i_j$ in $I]$."

Setting these two possibilities aside for a moment, recall Goodin's articulation of the background assumption interpretation: "If any one of the conditions presupposed by ideal theory is missing, then the Theory of Second-Best warns that we might... need to make systematic alterations right across the board in the prescriptions of ideal theory". Since the "conditions" discussed in the antecedent are those "*presupposed by* ideal theory", Goodin is clearly referring to ideal theorists' assumptions about the circumstances in which the ideal state of affairs (institutional scheme, distributive profile, or whatever) is to be realized — for example, sufficient material resources, few barriers to collective action, and so on. This is akin to the set of modeling assumptions deployed by economists (absence of monopolies, externalities, and transaction costs, for instance). Put simply, Goodin's interpretation of the antecedent refers to the *input* of a theoretical exercise: "If any one of $[i_1, \ldots, i_n]$ is missing..." (whence the moniker "background assumptions interpretation"). Equally clearly, Goodin's interpretation of the theorem's consequent refers to the *output* of ideal theory — its "prescriptions" or normative principles. These

15 Throughout the paper, the variable $j = 1, \ldots, n$ denotes an arbitrary index for members of a set.

are akin to the constitutive conditions for a Pareto efficient allocation, stated above. Thus, stating the background assumptions interpretation more precisely:

> *The Theorem of Second Best — BA.* If an arbitrary i in I fails to obtain,[16] then, for all j, the best state of affairs under the circumstances does not necessarily satisfy o_j in O.

More informally: if some background condition assumed by a theorist's model of the ideal state of affairs (e.g., institutional scheme, social practice, or whatever) fails to obtain, then we should not necessarily aim to satisfy the normative principles derived from an analysis of the model (assuming we should aim to realize the normatively best state of affairs under any given circumstance). So if Rawls's assumption that, say, people have a "sense of justice" fails to obtain, then we should not necessarily aim to satisfy Rawls's principles of justice.

The first thing to note about *The Theorem of Second Best — BA* is that it implies that L&L's use of "optimum conditions" is equivocal. Adhering to a principle of interpretive charity, this is enough to impugn the adequacy of this interpretation of L&L's theorem. To avoid the implied equivocation, we can hold the interpretation of the antecedent fixed and change the interpretation of the consequent or *vice versa*.

Holding the interpretation of the antecedent fixed yields *The Theorem of Second Best — Input*. This theorem might be true; but it fails to yield any worthwhile insight in the present context. To see why this is so, consider Rawls's theory of justice in the light of *Input*. As noted above, Rawls assumes (among other things) that individuals have a "sense of justice", that society's stock of material resources is sufficient to provide each citizen with a reasonably extensive set of basic liberties, and that there are no cross-border transactions among societies. Now suppose that, in fact, citizens do not typically possess a sense of justice in Rawls's sense; so one of Rawls's assumptions about the background conditions fails to obtain. We learn from *Input* that the (normatively) best state of affairs given that citizens do not typically possess a sense of justice is not necessarily one in which society is closed to cross-border transactions or has enough resources to protect basic liberties equally. This might be true — but it's beside the point of the theoretical exercise. In the first place, this putative "insight" presupposes that the best state of affairs *given that citizens have a sense of justice* is one in which society is closed to cross-border

16 Note that, as Lipsey (2007) clearly acknowledges, second best situations need not arise from constraints that are, strictly speaking, impossible to overcome; they might arise from policy-created constraints, which in some sense *can* be overcome.

transactions and has sufficient resources to protect basic liberties.[17] But Rawls suggests no such thing; certainly, he never defends such a claim. Rawls's assumptions about background social conditions are not meant to characterize some ideal state of affairs (even though he does suggest at various places that he means to model conditions that are "favorable" for the realization of justice). Rather, his assumptions are meant to isolate a particular social problem for further analysis — namely, the problem of how to distribute the benefits and costs of social cooperation among people who are generally willing and able to treat their fellow cooperators fairly. Relatedly, the putative "insight" gleaned from *Input* fails to appreciate that the point of Rawls's theoretical exercise is to analyze the constitutive features of the *ideal solution to the particular social problem* modeled by his assumptions about background social conditions. He does not aim to model the ideal state of affairs *tout court*. At any rate, if he meant to pursue the latter, he would have had to argue for the following claim: given that people generally possess a sense of justice, a state of affairs in which society is closed to cross-border interactions or experiences a moderate measure of resource scarcity is better (from a moral standpoint) than a state of affairs in which societies interact with each other on just terms or in which material resources are abundant. Yet he never attempts to argue for such a (patently implausible) claim.

Compare the lesson gleaned from *Input* with that delivered by *Output*, the only other non-equivocal interpretation. *Output* presupposes that the output of Rawls's theoretical exercise — his principles of justice — are meant to characterize certain constitutive features of the best state of affairs given the background social conditions he assumes. Given this, *Output* tells us the following: given that one of Rawls's principles of justice remains unsatisfied (the difference principle, say), the best state of affairs does not necessarily satisfy Rawls's principles of justice. This result is at least *prima facie* insightful, given the point of Rawls's theoretical exercise — again, to characterize an ideal solution to the problem of distributive justice among cooperators who are willing and able to treat each other fairly.

So the only plausible non-equivocal interpretation of the second best theorem is the *Output* version above. But, contrary to the background assumptions interpretation, *The Theorem of Second Best — Output* interprets "optimum conditions" in the antecedent as

17 In general, the second best theorem is a result pertaining to the circumstance in which some set of conditions understood to constitute an ideal state of affairs fails to be fully satisfied (this is common across all interpretations of the theorem). Accordingly, *Input* presupposes that the background conditions $i_1, \ldots, i_n$ assumed by a model jointly constitute the best state of affairs. Hence, it presupposes that the ideal state of affairs given i_j is one in which $i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_n$ obtain.

referring to the constitutive conditions for achieving Pareto efficiency (O), not as referring to the background conditions assumed by abstract economic models (I).

The foregoing remarks are only suggestive. A careful look at the mathematical details of L&L's proof is sufficient to decide the issue against the background assumptions interpretation: L&L's use of "optimum conditions" does not refer to the background conditions assumed by economic models (I) but to the constitutive conditions for Pareto efficiency derived from analyzing those models (O). We need not discuss these mathematical details here. The pivotal insight that arises from examining L&L's proof is this: a second best situation arises from a barrier to satisfying one of the derived constitutive conditions without regard for the cause of this barrier — in particular, whether or not it arises because one of the background conditions assumed by the initial specification of the model fails to obtain. (But the reader need not take my word for it; I rehearse the relevant details in an appendix below.) In terms of political theory, L&L's theorem applies when there arises a barrier to satisfying one of the normative principles derived from an analysis of a model of the ideal state of affairs (e.g., an institutional scheme, a social practice, a situation involving interpersonal transactions). Stating the antecedent strictly using the above terms, the theorem says:

> *The Theorem of Second Best — Strict.* If an arbitrary o_j in O remains unsatisfied, then, for all $k \neq j$, the best outcome under the circumstance does not necessarily satisfy o_k in O.[18]

This is just a re-articulation of *The Theorem of Second Best — Output*, stated above. The background assumptions interpretation of the theorem is mistaken.

Consider the application of the second best theorem to Rawls's theory of justice to illustrate the contrast between the two interpretations at issue. According to the mistaken background assumptions interpretation, L&L's theorem implies that we should not necessarily aim to satisfy Rawls's principles of justice if, for instance, the assumption of a closed society (or the assumption of sufficient material resources or whatever) fails to obtain in our world. In contrast, the correct "strict" interpretation of the theorem implies that we should not necessarily aim to satisfy Rawls's principles of justice *so long as at least one of those principles remains unsatisfied.* For instance, if Rawls's equal basic liberties principle remains unsatisfied, then (by the theorem of second best) we should not necessarily aim to satisfy the difference principle.[19]

18 The variable $k = 1, \ldots, n$, like j, denotes an arbitrary index for members of a set.

19 This example might already bring to light the divergent implications of the two interpretations for

One might object that we ought not interpret L&L's theorem as strictly as I have here. L&L's formalization of the second best problem merely illustrates a range of second best problems. After all, L&L's formulation only attends to "distortion[s] between price and marginal cost in some market(s)" and the "implications this ha[s] for pricing rules that ought to be followed by the other, non-distorted, sectors". But, nowadays, economists acknowledge that "distortions can arise for a variety of reasons" and that second best theory applies beyond pricing rules in non-distorted sectors of the economy (see Boadway, 1997, 3–4). Even Lipsey (2007) acknowledges many "sources of divergence", that is, many barriers to the realization of a Pareto optimal outcome. Hence, any theoretical exercise for which the circumstances assumed to obtain in the ideal state of affairs remain unsatisfied should be regarded as a second best problem.[20]

It is true that there are numerous barriers to realizing a Pareto efficient allocation, many of which arise due to the violation of economists' modeling assumptions: the existence of monopolies, incomplete markets, consumption externalities, or imperfect information. Hence, there are as many second best problems as there are barriers to realizing an efficient allocation. But this objection misses the mark. Boadway's and Lipsey's multifarious "sources" of distortion or divergence refer to different obstacles to fulfilling one of the constitutive conditions for a Pareto efficient allocation, that is, to satisfying one (or more) of $o_1, \ldots, o_n$. Oftentimes, the "source" in question might be a failure of some i in I to obtain. But it need not be. Moreover, the failure of some i to obtain need not entail a barrier to satisfying some o in O. In principle, at least, there are ways to achieve an efficient allocation despite the failure of some circumstance required for a competitive market equilibrium. After all, the first fundamental theorem of welfare economics states that a perfectly competitive market system is only a *sufficient* condition for Pareto efficiency (see Ng, 2004, ch. 2). This point gains strength when we shift to normative political theory, where there are no theorems demonstrating any analytical link between the circumstances presupposed by a model of some state of affairs (I) and the normative principles we derive from an analysis of the model (O), as the two fundamental welfare theorems do. Hence, the theorem of second best does not apply whenever *any* aspect of the ideal does not obtain. It applies specifically whenever at least one of the *derived* (rather than assumed) constitutive conditions for an ideal state of affairs remains unsatisfied.

ongoing methodological disputes among political philosophers. I discuss these implications in the next section.

20 This objection comes from an anonymous reviewer.

5. IMPLICATIONS FOR IDEAL THEORY

The discussion in the preceding section is not mere pedantry. How we interpret the theorem's antecedent is consequential for wielding it against the widely-held Target View (introduced above). Consider the argument against the Target View licensed by the background assumptions interpretation of the theorem (cf. Brennan and Pettit, 2005; Goodin, 2012; Heath, 2004):

(1) *Characterization of Ideal Theory.* Let $I = \{i_1, \ldots, i_n\}$ be the set of circumstances an ideal theorist (e.g., Rawls) assumes when specifying $O = \{o_1, \ldots, o_n\}$, the set of ideal normative principles (e.g., Rawls's two principles). At least one i in I will fail to obtain.

(2) *The Theorem of Second Best — BA.* If an arbitrary i in I fails to obtain, then, for all j, we should not necessarily aim to satisfy o_j in O.

(3) Therefore, for all j, we should not necessarily aim to satisfy o_j in O in the actual world.

This argument delivers a powerful challenge to the Target View because it need only appeal to a relatively uncontroversial characterization of ideal theory and an (allegedly) deductively proven theorem. But, as shown in the previous section, L&L do not prove premise (2). Moreover, this argument ceases to be valid if we insert the strict version of the theorem in place of (2):

(2′) *The Theorem of Second Best — Strict.* If an arbitrary o_j in O remains unsatisfied, then, for all $k \neq j$, we should not necessarily aim to satisfy o_k in O.

Premise (2) might be true, in which case, the conventional argument might nonetheless be sound. But demonstrating this requires an argument for (2), which philosophers have declined to provide since they typically defer to L&L at this point. Moreover, intuition and conventional wisdom seem to oppose (2). Many philosophers have argued that we should aim to satisfy ideal normative principles despite the failure of ideal circumstances to obtain in the actual world (see, e.g., Lawford-Smith, 2010; Mason, 2004; Simmons, 2010; Valentini, 2009). Conventional second best-based challenges to the Target View have yet to be vindicated.

A valid argument that turns on the second best theorem must take a pessimistic stance on the prospects for fulfilling ideal principles in the actual world:

(1′) Let $O = \{o_1, \ldots, o_n\}$ be the set of normative principles derived from ideal theory. There is at least one o_j in O that will remain unsatisfied in the actual world.

(2′) *The Theorem of Second Best — Strict.* If an arbitrary o_j in O remains unsatisfied, then, for all $k \neq j$, we should not necessarily aim to satisfy o_k in O.

(3′) Therefore, for all $k \neq j$, we should not necessarily aim to satisfy o_k in O in the actual world.

Premise (1′) is controversial, to say the least. Whether ideal principles will remain unsatisfied in the actual world and, hence, fail to present appropriate normative targets in the actual world is precisely what's at issue between ideal theory skeptics and proponents. There might be compelling evidence that some particular set of ideal principles (e.g., Rawls's two principles[21]) will remain unsatisfied in the actual world. But there does not seem to be any evidence to support general pessimism about satisfying ideal normative principles in the actual world. Indeed, some (myself included) might think the issue cannot, in principle, be resolved (Wiens, forthcoming*b*). A mere conjecture that fulfilling an ideal principle might not be satisfied is surely not enough to trigger the second best theorem. At least, the mere suggestion of pessimism will be insufficient to persuade many philosophers, most of whom think that the burden of proof surely rests with the skeptic.

So critics of the Target View have much work to do. On the one hand, they typically focus on the question of whether we should aim to satisfy normative principles specified under assumptions that fail to obtain in our world, occasionally appealing to the theory of second best to bolster their skepticism. But, contrary to popular belief, L&L's theorem does not address this specific issue; L&L do not prove premise (2) in the conventional argument. The theory of second best speaks to a distinct (although related) issue, namely, whether we should aim to approximate what remains of an ideal theory so long as one of the ideal principles remains unsatisfied. Hence, if the conventional argument is to pose a challenge for the Target View, premise (2) requires an argument yet to be given.

On the other hand, the argument deploying the strict version of the theorem does potentially pose a powerful challenge to the Target View insofar as ideal theory generally yields normative principles that will remain unsatisfied in the actual world. But whether ideal theory generally yields principles that will remain unsatisfied is highly controversial; as yet, we lack a compelling argument to support premise (1′). Worse (for Target View critics), it seems nigh impossible to demonstrate premise (1′) persuasively, short of

21 There are plenty of arguments meant to show that implementing Rawls's principles in our circumstances is *undesirable*; see, e.g., Farrelly (2007); Mills (2005). Few have discussed the *likelihood* of their fulfillment.

actually trying to implement a sufficiently wide range of ideal principles and failing to successfully do so. Thus, correctly interpreted, the theory of second best provides little reason to reject the general view that political ideals can present appropriate targets for real world reform efforts.

But the theory of second best brings cold comfort to ideal theorists. In the face of pessimism about the prospects for realizing a political ideal, ideal theorists often reply that the ideal presents an appropriate target for reform even if we know that we cannot ultimately realize the ideal in the actual world. This is so because, even in the pessimistic case, we do best (from the standpoint of justice) by realizing a state of affairs that satisfies as many ideal principles as possible or satisfies them to the greatest extent possible. Put differently, whatever the prospects for realizing a particular political ideal, it serves as an appropriate reform target because we should try to realize a state of affairs that approximates the ideal as far as possible (see, e.g., Christiano and Braynen 2008; Gilabert 2012*b*, p. 243; Valentini 2012*b*, p. 42). The second best theorem straightforwardly undermines this "approximation" reply. The theorem says that we can have no reasonable expectation that the best state of affairs short of fully realizing an ideal is a state that approximates an ideal as far as possible. Thus, absent credible evidence that the ideal is sufficiently likely to be realized, we have no reason to expect that a political ideal presents an appropriate target for real world reform efforts.[22]

In view of this anti-approximation warning, those who wish to uphold a particular ideal as an appropriate reform target must adopt one of two strategies, either of which takes the ideal theorist beyond the limits of conventional ideal theory. First, one can *circumvent* the second best theorem by presenting credible evidence for optimism about the prospects of realizing the proposed ideal. This requires the ideal theorist to go beyond simply analyzing the constitutive features of the ideal state of affairs. To provide a credible basis for optimism, one must at least analyze the causal mechanisms that generate the status quo and analyze whether any of the available interventions are sufficiently likely to realize the ideal (Wiens, 2013).[23] Hence, ideal theorists who wish to uphold their proposed ideals as reform targets cannot be satisfied with the standard optimistic conjectures about the causal processes that could — in some possible world — realize the proposed ideal.

Second, in lieu of presenting credible evidence that the proposed ideal is sufficiently likely to be realized, an ideal theorist might *counter* the second best theorem by demon-

22 Cf. L&L's "negative corollary": if at least one of the optimum conditions remains unfulfilled, then we cannot determine a priori whether movement toward greater fulfillment of the unconstrained optimum conditions constitutes an improvement over the status quo (L&L, 12).

23 I leave the sufficiency threshold unspecified to avoid any commitments about how heavy the burden of proof is.

strating that, among the feasible alternatives,[24] the best state of affairs (from the standpoint of justice) is the state most likely to emerge from efforts to realize the ideal as closely as possible. This, too, requires the ideal theorist to go beyond simply analyzing the constitutive features of the ideal. At a minimum, vindicating this comparative claim requires one to (1) estimate the state of affairs most likely to emerge from efforts to realize the ideal as closely as possible, and (2) compare this state with feasible alternatives from the standpoint of justice.[25] Notice that such a comparative exercise makes the conventional practice of analyzing ideals redundant for the practical purpose of figuring out which states of affairs we should aim to realize. Learning that a particular state of affairs — let's call it s^* — is the best feasible state of affairs from the standpoint of justice is sufficient to show that we should aim to realize it. We gain no additional reason to pursue s^* from the knowledge that s^* satisfies the constitutive features of the ideal as closely as possible (Wiens, forthcoming*a*).[26]

In sum: A correct understanding of the theory of second best bears two lessons for ongoing methodological disputes among political philosophers. First, the theory of second best fails to provide a compelling reason to reject the Target View in general; normative theorists need not abandon the business of proposing political ideals as targets for real world reform efforts. But, second, the anti-approximation warning posed by the theory of second best implies that normative theorists can uphold particular ideals as reform targets only if they undertake the kinds of causal and comparative analyses typically neglected by conventional ideal theories.

6. APPENDIX

My aim in this appendix is to make apparent why the background assumptions interpretation of the theorem is misguided. To do this, I walk through the key details of L&L's proof, presented in sec. 7 of the original paper.[27]

24 Nothing here turns on how we analyze the concept of feasibility. For the curious, I present my preferred analysis in Wiens (forthcoming*b*).

25 Actually, one must estimate the probability distribution over states of affairs that emerges from efforts to realize the ideal as closely as possible and compare this probability distribution with those that emerge from feasible alternative courses of action. As nothing I say here turns on this complication, I set it aside.

26 One might counter that we need a characterization of the ideal to provide the metric by which we judge feasible states of affairs from the standpoint of justice. I deny this "benchmark view" in Wiens (forthcoming*a*).

27 There are some typographical errors in the original proof, which are corrected in Lipsey and Lancaster (1997). Ng (2004, ch. 9) presents a more accessible proof, including a helpful graphical illustration of the theorem's key implication.

L&L's proof assumes a generic differentiable objective function of n variables, $F(x_1,\ldots,x_n)$. This serves to rank states of affairs as a function of the characteristics measured by $x_1,\ldots,x_n$. This function is to be optimized subject to a differentiable constraint on the variables, $G(x_1,\ldots,x_n)=0$. This serves to specify the background conditions, which define the set of feasible states of affairs. If we select an arbitrary variable, say x_n, to serve as the standard of relative value (the "numeraire"), then, using the Lagrange method, we know that the (constrained) optimum obtains just in case

$$\frac{F_i}{F_n}=\frac{G_i}{G_n} \text{ for all } i=1,\ldots,n-1, \tag{1}$$

where $F_i=\frac{\partial F}{\partial x_i}$ is the partial derivative of F with respect to x_i and $G_i=\frac{\partial G}{\partial x_i}$ is the partial derivative of G with respect to x_i.

Welfare economists interpret F as a social welfare function and G as a production constraint. Hence, the optimum conditions depicted in (1) are interpreted as the third Paretian condition above (section 2): the common marginal rate at which consumers substitute good x_i for x_n must be equal to the marginal rate at which the economy can redirect production of x_i into production of x_n. But it's important that the theorem's application does not turn on this interpretation. Since F and G are both generic functions, these derived conditions more generally represent the optimal relationships among the objective function's constituent variables. What's important here is that $x_1,\ldots,x_n$ represent evaluative criteria, F represents an evaluative standard (i.e., the balance of evaluative criteria), G represents a specification of the background conditions (i.e., constraints on the realization of the evaluative criteria), and the set of equations in (1) represent the optimum conditions derived from F and G, the constitutive conditions for an optimal state of affairs.

The next step in the proof introduces a second best situation, modeled as a constraint k that prevents an arbitrary optimum condition from being satisfied. Since nothing hangs on the choice of condition, we can suppose

$$\frac{F_1}{F_n}=k\frac{G_1}{G_n} \text{ for some } k\neq 1. \tag{2}$$

Again using the Lagrange method, the function to be optimized given this new constraint is

$$F(x_1,\ldots,x_n)-\lambda G(x_1,\ldots,x_n)-\mu\left(\frac{F_1}{F_n}-k\frac{G_1}{G_n}\right). \tag{3}$$

This is just the original optimization problem with the unsatisfied optimum condition appearing as a new constraint. The proof proceeds to derive the new optimum conditions for an efficient outcome given the introduced constraint — a "second best optimum" — showing that the second best optimum conditions are not identical to the first best conditions given in (1). We can set this last part of the proof aside here. The key point is that L&L's use of "Paretian optimum conditions" in the antecedent of the theorem denotes the first-order optimum conditions derived from the objective and constraint functions; it does not refer to the circumstances presupposed by the original constraint function. Indeed, the new optimization problem need not alter G in any way. That is, the new optimization problem can retain all the same assumptions about background conditions. It must only add the fact that one of the optimum conditions is unsatisfied (whatever the explanation). Thus, it is a mistake to claim — as the background assumptions interpretation does — that the theorem applies when one of the background conditions assumed by the original constraint function is not satisfied in the actual world.

REFERENCES

Blackorby, Charles, Russell Davidson and William Schworm. 1991. "The Validity of Piecemeal Second-Best Policy." *Journal of Public Economics* 46:267–290.

Boadway, Robin. 1997. The Role of Second-Best Theory in Public Economics. In *Trade, Technology and Economics: Essays in Honour of Richard G. Lipsey*, ed. B.C. Eaton and R.G. Harris. Cheltenham, UK: Edward Elger pp. 3–25.

Brennan, Geoffrey and Philip Pettit. 2005. The Feasibility Issue. In *Oxford Handbook of Contemporary Philosophy*, ed. Frank Jackson and Michael Smith. Oxford: Oxford University Press.

Christiano, Thomas and Will Braynen. 2008. "Inequality, Injustice, and Levelling Down." *Ratio* 21(4):392–420.

Cohen, G. A. 2008. *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press.

Coram, Bruce Talbot. 1996. Second Best Theories and the Implications for Institutional Design. In *The Theory of Institutional Design*, ed. Robert E. Goodin. New York: Cambridge University Press.

Davis, Otto A. and Andrew B. Whinston. 1965. "Welfare Economics and the Theory of Second Best." *The Review of Economics and Statistics* 32(4):1–14.

Estlund, David. 2008. *Democratic Authority*. Princeton and Oxford: Princeton University Press.

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

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.

Goodin, Robert E. 1995. "Political Ideals and Political Practice." *British Journal of Political Science* 25(1):37–56.

Goodin, Robert E. 2012. The Bioethics of Second Best. In *Global Justice and Bioethics*, ed. Joseph Millum and Ezekiel J. Emanuel. New York: Oxford.

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

Heath, Joseph. 2004. "Dworkin's Auction." *Politics, Philosophy, and Economics* 3(3):313–335.

Lawford-Smith, Holly. 2010. "Ideal Theory — A Reply to Valentini." *Journal of Political Philosophy* 18(3):357–368.

Lipsey, R. G. and Kelvin Lancaster. 1997. The General Theory of Second Best. In *The Selected Essays of Richard G. Lipsey, Volume 1: Microeconomics, Growth and Political Economy*, ed. Richard G. Lipsey. Cheltenham: Edward Elger.

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

Lipsey, Richard G. 2007. "Reflections on the General Theory of Second Best at its Golden Jubilee." *International Tax and Public Finance* 14:349–364.

Mason, Andrew. 2004. "Just Constraints." *British Journal of Political Science* 34:251–268.

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

Ng, Yew-Kwang. 2004. *Welfare Economics: Towards a More Complete Analysis*. New York: Palgrave Macmillan.

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

Räikkä, Juha. 2000. "The Problem of the Second Best: Conceptual Issues." *Utilitas* 12(2):204–218.

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

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

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

Valentini, Laura. 2012*b*. *Justice in a Globalized World: A Normative Framework*. New York: Oxford 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*. "Against Ideal Guidance." *Journal of Politics* .

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