https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

ON THE VERY IDEA OF IDEAL THEORY IN POLITICAL PHILOSOPHY*

By Alexander Rosenberg

Abstract: *The essay agues that there is little scope for ideal theory in political philosophy, even under Rawls's conception of its aims. It begins by identifying features of a standard example of ideal theory in physics — the ideal gas law, PV=NRT and draws attention to the lack of these features in Rawls's derivation of the principles of justice from the original position. A. John Simmons's defense of ideal theory against criticisms of Amartya Sen is examined, as are further criticisms of both by David Schmidtz. The essay goes on to develop a conception of the domain of social relations to be characterized by justice that suggests that as a moving target it makes ideal theory otiose. Examination of Rawls's later views substantiate the conclusion that ideal theory as propounded in A Theory of Justice is a mistaken starting point in the enterprise of political philosophy. Differences between the domains of ideal theory in mathematics, physics, and economics on the one hand, and political philosophy on the other, reinforce this conclusion.*

KEY WORDS: ideal model, favorable circumstances, noncompliance, deformable elastic surface, social reflexivity, pure theory

I. Introduction

The ideal/nonideal distinction in political philosophy and the debate about it emerges from Rawls's *A Theory of Justice* (with brief amplification in *The Law of Peoples*).[1] In this essay, I begin by criticizing the distinction as Rawls draws it, and as it has been defended by able expounders of Rawls's doctrine. Then I ask whether there is a role for ideal theory — Rawlsian or other kinds — in political philosophy. As a prolegomenon I briefly consider one well-known successful ideal theory in science as an inspiration for Rawls.

Perhaps the scientific model for ideal theory that Rawls was thinking of, or that we might think of to try to make sense of his appeal to the notion is the Ideal Gas Law,

$$PV = nRT$$

where P is pressure, V volume, T temperature, n is the number of moles of the gas, and R is a constant composed of Avogadro's and Boltzmann's

[*] The author expresses his gratitude to David Schmidtz, an anonymous referee, and the other contributors to this volume.

[1] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971); *The Law of Peoples* (Cambridge, MA: Harvard University Press, 2001)

doi:10.1017/S0265052516000376

constants. What makes this law ideal is not the deviation of reality from the values it asserts to obtain. In fact, at all but the highest pressures and smallest volumes, the law provides very accurate predictions of real values. It is a real gas law, for all gases and most values of pressure, temperature, and volume known to nineteenth-century physics. What makes the law *ideal* is its derivation in the kinetic theory of gases. Indeed, prior to this derivation, it was expressed in a set of laws, none of which were considered ideal: Boyle's gas law, Charles's gas law, Guy-Lussac's gas law. The equation came to be called the ideal gas law owing to the fact that it was derived in the kinetic theory of gases from assumptions in which we had great confidence (that molecules, if they exist at all, obey Newton's laws) and assumptions about molecules that were known to be false, but in this case harmless idealizations: that gas molecules are point masses — take up no volume despite their mass, and that gas molecules are perfectly elastic in collision, contrary to Newton's laws and the laws of thermodynamics.

Of course, as experimentalists were able to increase the pressure and decrease the volume of gases, the relationship between these two variables and temperature began to deviate from the ideal gas law. The nineteenth and twentieth centuries saw a series of additions to this equation that increased its range of accurate prediction by adding variables that reflected the size of gas molecules, the degree of compressibility of a gas, and the intermolecular forces that made their collisions imperfectly elastic. The result was a succession of new gas models, each with a characteristic gas law equation.

Something to notice here: the original ideal gas law was highly successful in predicting real values of any one of its three variables from measurement of the other two. Confidence in it did not depend on a derivation of it that deliberately ignored forces independently known to obtain. For that reason, we knew from the outset that the idealizations made were harmless in a wide range of circumstances.

How much of a model is there here for ideal theory in Rawls's hands? Setting aside the difference between a positive theory and a normative theory, some similarities are obvious. In *A Theory of Justice*, a simple result is derived from assumptions, some of which Rawls was very confident in — the employment of a maximin strategy — and some of which we know not to universally obtain — the veil of ignorance, perfect compliance, and moderate scarcity. So much also holds for the ideal gas law.

But the disanalogies are glaring. To begin with, the ideal gas law secured a wide acceptance long before it began to be treated as an idealization. First formulated in 1834, it was only derived from idealizing assumptions in thermodynamics in 1856. No such prior general acceptance characterizes the principles of justice that Rawls derives in the original position. Moreover, the idealizations that were invoked to explain the ideal gas law were held to be harmless, first because the law derived from them so neatly and directly was well confirmed, so there was no symptom of harm to pin on the idealizations; second, because the same sort of assumptions

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

made elsewhere in mechanics — for example, in the role given to centers of mass — had independently been shown to be harmless; and third, because when made, these ideal assumptions could be nicely combined with the most well-established assumptions of physics — Newton's laws — to explain the gas law itself. None of these facts seem to have a parallel in the case of Rawls's derivation of the principles of justice from the application of a maximin strategy behind the veil of ignorance in the original position. The disanalogies vastly outweigh the analogies.

Accordingly in most of what follows, I am going completely to set aside any thoughts about Rawlsian ideal theory as even a dim echo of the best candidate for a similar ideal theory approach in natural science. Once I have traced how unsuitable the idea of ideal theory is in political philosophy altogether, I will return to a comparison of Rawlsian ideal theory with some other examples of ideal theory — in mathematics, in physics, and in economics. My conclusions about its prospects will be only a little more optimistic than the ideal gas law comparison suggests.

II. Ideal Theory, Noncompliance and Favorable Circumstances

According to Rawls, "ideal theory" addresses the question of "what a perfectly just society would be like."[2] Therefore, Rawls immediately infers that the ideal theory of justice may exclude two features of real life: noncompliance and unfavorable conditions. He does not argue against or consider the claim that a perfectly just society needs to be one that is perfectly just in its treatment of noncompliance and scarcity, even though these are both factual assumptions one can safely make about all real societies. Many such factual assumptions are built into ideal theory: its strictures on just institutions must be "realistically practicable." These constraints take into account the general facts of moral psychology: ideal theory does not require moral heroism. The thought experiment in which, according to Rawls, rational agents will concur in his theory of justice must proceed on the assumption of *favorable conditions*. Ideal theory is a set of claims about what institutional arrangements are just when scarcities are not so severe that a "constitutional regime," in particular a democratic one, is possible. This assumption mirrors one David Hume made in his *Enquiry Concerning the Principles of Morals*,[3] arguing that justice emerges under conditions of "moderate scarcity." There is a pretty compelling argument in Hume for assuming something like moderate scarcity in political philosophy: without scarcity there is little need to craft just institutions.

Though Rawls does not seem to notice it in *A Theory of Justice*, these two conditions — compliance and favorable conditions — are not unrelated.

[2] Rawls, *Theory of Justice*, 8.

[3] David Hume, *An Enquiry Concerning the Principles of Morals*, [1751].

In fact, their connection makes it difficult to combine both of them in one ideal theory.

The inclusion of favorable or better conditions, moderate scarcity, and the exclusion of noncompliance from ideal theory is *prima facie* odd. *Ceteris paribus,* the more favorable the conditions in a society are in general, the less noncompliance, and *vice versa*, the greater the scarcity, the more noncompliance. Hume recognized this relationship clearly enough:

> Let us suppose that nature has bestowed on the human race such profuse abundance of all external conveniences, that, without any uncertainty in the event, without any care or industry on our part, every individual finds himself fully provided with whatever his most voracious appetites can want, or luxurious imagination wish or desire. . . . No laborious occupation required: no tillage: no navigation. Music, poetry, and contemplation form his sole business: conversation, mirth, and friendship his sole amusement. It seems evident that, in such a happy state, every other social virtue would flourish, and receive tenfold increase; *but the cautious, jealous virtue of justice would never once have been dreamed of* . . . .
>
> To make this truth more evident, let us reverse the foregoing suppositions; and carrying everything to the opposite extreme, consider what would be the effect of these new situations. Suppose a society to fall into such want of all common necessaries, that the utmost frugality and industry cannot preserve the greater number from perishing, and the whole from extreme misery; *it will readily, I believe, be admitted, that the strict laws of justice are suspended, in such a pressing emergence, and give place to the stronger motives of necessity and self-preservation*.[4]

Given the close connection between compliance and favorable conditions, it is odd that one should be included and the other excluded from ideal theory. "Favorable" cannot mean "abundance" and so it cannot, without argument, exclude the possibility of some noncompliance. Are Rawls's reasons for doing so compelling?

A. John Simmon develops Rawls's rather underdeveloped argument for doing so:

> First . . . if we compare the operation of societies ordered by competing principles of justice while assuming strict compliance with those principles, the different effects we observe can reasonably be taken to be wholly the responsibility of the different ordering principles themselves. So our comparison turns out to be quite strictly a comparison only of the principles of justice.[5]

[4] Ibid., section III. Emphasis added.

[5] A. John Simmons, "Ideal and Nonideal Theory," *Philosophy and Public Affairs* 38, no. 1 (2010): 5–36, at p. 8.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

These reasons seem to be faithful to Rawls's brief comments in *A Theory of Justice*. But as reasons for both Simmons's and Rawls's exclusion of noncompliance from ideal theory Simmons's reasons are unsatisfying. There is, to begin with, the presumption that since circumstances can vary all the way from abundance to favorable to extreme privation, it seems arbitrary to stipulate favorable but not abundant conditions when seeking ideal principles of justice, but not also to stipulate some noncompliance. What is more, justice includes justice toward noncompliant parties: whether punishment is just, what sorts of punishments are just, whether justice requires a role in punishment for immediate victims of noncompliance, or compensation by the noncompliant, and so on. Even Rawls recognized this, albeit much later in *A Theory of Justice*: "[W]e need an account of penal sanction however limited even for ideal theory."[6]

Simmons tells us that if we assume a "normal" level of noncompliance "we will likely find both that our evaluations yield quite indeterminate results and that the results depend on more than simply the different ordering effects of the principles being compared."[7] But, first of all, assuming away noncompliance will itself make the results indeterminate insofar as we require just responses to noncompliance. Second, it does not seem to add much to the burdens of the original position to ask bargainers behind the veil of ignorance to consider a rational response to at least some noncompliance with the Rawlsian package to which they agree. Simmons's concern will seem arbitrary and abstract unless fleshed out by real examples. If, as Rawls acknowledges, parties to the bargain need to consider which principles of justice will generate their own support or lack stability, they are already contemplating the consequences of noncompliance. For this is what instability at least in part consists in. In fact, since Rawls makes the assumption that parties to the bargain understand the general facts of moral psychology and are not themselves moral heroes, they must have concerns behind the veil of ignorance about the just treatment of inevitable noncompliance. Or at any rate, Rawls needs a more compelling reason than a desire to simplify his problem by excluding noncompliance. Abundance would simplify his problem even more, but Rawls does not make that assumption.

Simmons also tells us on behalf of Rawls that "ideal theory cannot set 'partial targets' until it first determines that hitting those targets will be consistent with all other aspects of societal justice."[8] This limitation of ideal theory is yet another reason that it — ideal theory, and

[6] Rawls, *Theory of Justice*, 241.

[7] It is worth noting that the degree of complication that is introduced by what Simmons calls a "normal level of noncompliance" depends largely on the stringency of the strictures of justice that deem a level of noncompliance high, low, or "normal." Noncompliance is not a dependent or endogenous variable within a theory of justice. I owe this observation to David Schmidtz.

[8] Simmons, "Ideal and Nonideal Theory," 22.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

not just nonideal theory — needs to include as part of its conception of justice how noncompliance is justly to be treated, on the threat, in Simmons's terms of "hitting . . . targets" that will not "be consistent with all other aspects of social justice."

A theory of justice can be excused from addressing requirements of justice under conditions of abundance and conditions of extreme privation. In the latter there is no noncompliance because justice does not come in to it. In the former, there is none of either. The question of justice arises in the middle, when there are both favorable enough circumstances and some noncompliance.[9]

II. Mount Everest, Sink-Holes in the Desert, and the Elastic Landscape of Justice

Rawls famously held that "until the ideal theory is identified . . . nonideal theory lacks an objective, an aim, by reference to which its queries can be answered." In Simmons's words: "Ideal theory [must] have priority over nonideal theory . . . to dive into nonideal theory without an ideal theory in hand is simply to dive blind, to allow irrational free rein to the mere conviction of injustice and to eagerness for change of any sort."[10]

Rawls's claim that ideal theory is required as a target for meliorative steps in the right direction has been famously challenged by Sen. Employing the metaphor of Everest, he argued that we don't require to know its height in order to compare the heights of lesser peaks. *Mutatis mutandis*, we don't need ideal justice as a standard by which to compare nonideal alternatives for greater or lesser degrees of justice.

The mountaineering metaphors Sen employs are joined to another set that is in some ways more apt: comparing works by Monet and Manet for excellence does not require taking sides on whether the Mona Lisa is the best among works of art. It is a weakness of the debate about ideal theory that it seems to have taken one metaphor — the Himalayan one — too seriously. Sen writes,

> There would be something deeply odd in a general belief that a comparison of any two alternatives cannot be sensibly made without a prior identification of a supreme alternative. There is no analytical connection at all. . . . A transcendental identification is . . . neither necessary nor sufficient for arriving at comparative judgments of justice.[11]

[9] In light of approaches to justice as a moral ideal such as those of Estlund and Cohen, one might ask whether strictures on justice to which everyone always complied would still include strictures on just punishment of noncompliance. I leave this question to exponents of such an idealistic approach.

[10] Ibid., 34.

[11] Amaryta Sen, *The Idea of Justice* (London: Penguin, 2009), 102.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

Simmons challenges Sen's claim against the indispensability of ideal theory in part by taking Sen's metaphor literally: "which of two 'smaller' peaks of justice is higher (or more just) is a judgment that matters conclusively only if they are both on equally feasible paths to the highest peak of perfect justice.[12] And in order to endorse a route to that highest peak, we certainly *do* need to know which one that highest peak is."[13] If Sen's metaphor is entirely apt, Simmons has a point, one well understood in evolutionary biology and other domains in which there are paths to local optima that turn out to be *cul-de-sacs*. Consider, for example, the confluence of the human esophagus and the larynx. This evolutionary imperfection is the result of two optimizing trajectories that together produced an imperfection that will never be corrected in the evolutionary history of our species[14] —too much developmental genetics would have to be unraveled. Of course, in evolutionary phylogenetics, we have excellent reason to take seriously the applicability of multiple local equilibria that make an approach to a more generally, optimal equilibrium difficult, and we have molecular genetics to establish that in many cases this optimal equilibrium is flatly impossible. In the human sciences we have almost no such reliable theory or powerful predictive tools.

David Schmidtz has cast doubt on Simmons's conclusions in part by rejecting the Sen-inspired Everest metaphor in favor of one he considers more apt. I shall argue that even Schmidtz's metaphor still obscures an important feature of human life that thoroughly undermines the claim that we need or can even have such a theory as an evaluative standard.

Schmidtz writes:

> If we take Sen's metaphor at face value, there is no question. Simmons is right. . . . The metaphor is Sen's. If it misleads as astute a critic as Simmons, Sen has only himself to blame. Sen scarcely gestures at an argument here. . . . I think what Sen needs to say is that the terrain's outstanding landmarks are injustices: pits in an otherwise featureless plane. Why don't we need to theorize about remote peaks? Answer: because they don't exist. Justice has no peak form. For thousands of years we postulated that it did, but we never had any reason, and we were wrong. There is no climbing to be done, no destination to seek, no problem to solve, unless people are in one of those pits. All we need to know about is the pits: what counts as being in, what counts as climbing out.[15]

[12] This claim is by itself debatable. If the Rawlsian "Everest" is unattainable, then neither of the "lesser peaks" of justice would be on a feasible path to it. Yet it would still be an important matter which is to be preferred in a theory of justice. I owe this point to David Schmidtz.

[13] Simmons, "Ideal and Nonideal Theory," 35.

[14] Indeed, one may speculate that the evolution of speech in *Hominins* took advantage of this regrettable imperfection. I owe this observation to David Schmidtz.

[15] David Schmidtz, "Ideal Theory: What It Is and What Ideally It Would Be," *Ethics* 121 (2011): 775–76.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

So, Schmidtz tells us that the landscape of justice is more like a desert pockmarked by pits than it is like a Himalayan mountain range.

Which metaphor is apt here may strike one as not a very important matter. But the exploitation of metaphor to deliver a theory and convey its force is by no means to be discouraged in science or political philosophy. On the contrary we have a great deal of evidence in the sciences to conclude that metaphors are not just effective. They are indispensible in expressing a theory; more than that, they may be cognitively unavoidable. Imagery, especially in the domain of theoretical physics, makes plain the effectiveness, indispensability, and even unavoidability of metaphor. The problem with Schmidtz's substitution of the desert landscape pockmarked with sinkholes for Sen's mountain range is that it still leaves out an essential feature of the landscape within which we seek just institutions and improvements in them. Improving on the metaphor, adding some features may blunt the force of Rawls's comeback that Schmidtz's dismissal of ideal theory, as a several-thousand-year mistake, is itself unsupported by any more than a metaphor.

A better (but still almost certainly misleading) metaphor for the landscape of justice is something like the deformable elastic sheet often used to illustrate general relativity. The standard example in physics is a trampoline on which a variety of balls are placed: a bowling ball, a billiard ball, a baseball, a squash ball, a Ping-Pong ball. Each makes a different indentation deforming the flat surface of the elastic sheet. The bowling ball's indentation is of course much greater than the others; the Ping-Pong ball's indentation is imperceptible. In simplifications of general relativity, the deformations represent the way in which mass curves space and produces the illusion of gravitational force. Now, as the balls move around the surface, they carry their indentations with them, and the heavier of them affect the paths of the lighter ones. If some balls are removed, there will be changes in the location of the others and (if total elasticity is finite) perhaps even changes in the amount of deformation produced by the remaining ones. We have a desert landscape with sinkholes that appear, disappear, change shape and location, over time, and influence one another's sizes and shapes as they do so.

The domain in which we seek enhancements in justice is more like this elastic sheet than Schmidtz's desert landscape with sinkholes or Sen's Himalayan mountain chain. The space with which the theory of justice deals is composed of human relations, relations of individuals to one another, between individuals and groups, and relations of groups to groups. These relations are characterized by packages of strategies —cooperative and competing ones, played by individuals and groups, played only once, or occasionally, or regularly, or for long periods, depending on their payoffs for players. Distinctive practices and institutions in societies consist in such packages of strategies.

The reason may be obvious to some students of social science. Human action, indeed human behavior, is highly reflexive. Among humans,

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

choices are almost always to some extent strategic, not parametric. People choose their behaviors to line up with or exploit other people's strategies. Individual behavioral strategies get packaged together into the various practices, institutions, and cultural artifacts that characterize social life. Social and political institutions are packages of strategies. These institutions exist for varying lengths: some like slavery or feudalism may last for a thousand years, others like fist bumping may last a few months. Some of them are designed — the U.S. Senate — others are constructed over time — the British House of Commons — others are composed of human strategies that exist not because they serve our purposes, recognized or not, but because they parasitize us — for example, tobacco smoking or foot binding. Human institutions, practices, and behaviors are all subject to shaping — indeed, subject to Darwinian selection, modification, adaptation — by the other packages of strategies that constitute their environments. So, like adaptations in the biological domain, they continually evolve, and like the most complicated of biological adaptations, their evolution is frequency dependent, co-adapted to persist in local equilibria, but continually "searching through design space" to find ways to take advantage of, exploit, outcompete, and sometimes temporarily cooperate with other packages of strategies.

Since which individual and group strategies are chosen often depends on which strategies are already in play, the space of human institutions and behaviors is a constantly shifting landscape in which there are few regularities that obtain long enough for policy planning to actually exploit in the design of institutions.

Two reflections of this shifting character of human social life are the difficulty of identifying unobtrusive or incentive-compatible measures of human behavior, and the ways in which regulatory institutions and regulated institutions engage in arms races. Sen in particular has noted that it is impossible to measure human capabilities in the way that means testing requires, without changing the expression of the capabilities that are measured. When it comes to distributive justice in the provision of scarce resources, there is a perpetual arms race between new measures of means testing and strategies of dissimulating actual capability.[16] The history of banking regulation in the United States and elsewhere reflects the same reflexive pattern in which banks and other financial institutions consistently seek new ways around regulation (including regulatory "capture"), and regulatory agencies add regulations that are always at least a step or two behind the regulated.

Onto this conception of human social relations and the institutions they create and constitute, we can impose a further highly variable

[16] Amartya Sen, *Development as Freedom* (New Haven, CT: Yale University Press, 1997), chap. 6.

topography, driven by considerations of justice. Schmidtz talks of sinkholes and Simmons of mountain peaks. If we take the metaphor of the elastic sheet seriously, then our evaluative or normative perspective will identify some regions as sinkholes of injustice — for example, slavery — and others as high ground of justice in distribution — say a national health service. Taking the metaphor seriously, however, will reveal that justice is very much a moveable feast or a moving target, or, better, a path along the elastic sheet that is in constant need of readjustment. The terrain of justice is a continually changing surface in which there are sinkholes and perhaps hilltops, but in which these sinkholes and hilltops grow and shrink continually, even move, fission, fuse, and most complicating of all, continually create new features — valleys, gullies, troughs, berms, bluffs, hills, mountains, cliffs, and so on, as they come and go. They do so because moving a large group out of one sinkhole will inevitably change the local shape of the elastic sheet in which other individuals or groups find themselves. Depending how deep the sinkhole, moving people out of it will result in substantial changes in their own strategies of interaction and in those of many other individuals and groups, sometimes producing new sinkholes into which individuals and groups have sunk where there were none before, or perhaps making significant improvements for everyone everywhere in those cases where enhancements in justice are non-zero sum in their effects. Consider how steps toward women's suffrage and African-American enfranchisement in the United States during the nineteenth century effected injustices on both women and former slaves.

The upshot of these pervasive facts about social life for our conception of the terrain of justice is clear. It is composed of no fixed chain of mountains in which the nearest obscure the tallest, nor a desert landscape with sinkholes located at fixed points that can be "filled in" without possibly widespread ramifications — erosion and even subsidence elsewhere in the desert.

If this metaphor is more nearly apt than the one Sen employs and Simmons seem to take so seriously, or the one that Schmidtz substitutes, the implications for ideal theory are pretty clear. It's not merely that ideal theory is neither necessary nor sufficient for identifying improvements that move us toward some maximally attainable level or quantity of institutional justice. There is indeed no permanent Everest in the actual landscape to begin with. At most there are multiple local promontories whose altitudes about the plain are hard accurately to measure. And just climbing out of one pit may eventually lead us to another, even worse pit of injustice. What is worse yet, climbing out may itself sometimes be part of the process that produces the deepest pit. Think of how some regard the harms to which capitalism's eclipse of feudalism led. And insofar as our powers to predict the ways the terrain of justice may change over time is very limited, and if these powers are not likely to improve, there is little scope for an ideal theory to serve as a standard of justice at all.[17]

For all its complexity, even the elastic sheet metaphor is overly simple. Justice really should be conceived as a hyperspace, whose dimensions are given by the many different kinds of consequences on individuals and groups that their movements and the movements of others in the space produce. There are both unintended and unforeseen consequences that need to be folded into any account of the impact on everyone (including everyone else) of moving people out of sinkholes or up to local optima.[18]

Schmidtz argues that theories are not sets of propositions, or arguments, that resist counterexamples; theorizing is not an attempt to lay out in conditional statements the necessary and sufficient conditions for the subject matter of theories. Theories, he writes, are maps. Presumably this is a claim about theories in political philosophy. It is not an unpopular view in the philosophy of science as well, or at least was not during a period in which instrumentalism about scientific theories was more fashionable. The map metaphor has even greater attractions in normative theory since it is supposed to be action guiding, unlike purely descriptive theory that by itself does not counsel or enjoin action. Maps, after all, are made for purposes. Schmidtz writes: "a map is not a truth maker, but a truth-tracker, at best providing useful but fallible guidance in navigating what is real, namely the terrain. . . . It is not only maps that are incomplete. The terrain being described (that is, justice itself) can be incomplete as well."[19]

The metaphor of a theory of justice as a map that is a work-in-progress is an apt one if we accept the notion that the terrain of justice has the features

[17] Is it enough of a vindication of ideal theory that it offer a standard of justice that applies "temporarily" or to a time slice "snapshot" of social relations and institutions, enabling us to identify the highest peak of justice on the current landscape, the one closest to fulfilling the standards of ideal theory, and perhaps also enabling us to temporarily prioritize the gravest injustices to rectify? The trouble with this rather modest ambition for ideal theory is that movements in directions toward and away from the temporary, perhaps transitory "locations" identified using the standard, will shift the landscape itself. I give some examples below.

It is true that justice-enhancing amelioration requires some kind of measure on the dimensions of the space. This by itself does not make ideal theory necessary or even feasible. Following G. A. Cohen, *Rescuing Justice and Equality* (Cambridge, MA: Harvard University Press, 2008), one might set a higher ambition for ideal theory, one that identifies the highest point or altitude in the space of justice. But if such a point is not attainable without, so to speak, rending the elastic fabric of society altogether, the approach is of little more than academic interest.

[18] At the risk of pushing the metaphor too far, we may initially think of the elastic surface proposed as having x,y coordinates reflecting social relations, and an orthogonal z-axis reflecting degrees of justice and injustice. I owe this observation to a referee. However, as the text notes, if justice is a multidimensional "quantity," we need to expand the space well beyond three dimensions to a "hyperspace."

[19] Schmidtz, "Ideal Theory: What It Is and What Ideally It Would Be," 775–76. Using a map for action-guiding purposes in the domain of justice complicates the elastic sheet metaphor further. The map cannot track fixed truths. The actions it guides, to move out of sinkholes or further up hillsides to locally more just outcomes, will also change the terrain on which the map is supposed to provide guidance. So, the normative map cannot just track a preexisting truth, as realism requires a theory/map to do in science.

described above. It is a continually changing surface composed of human practices and institutions whose topography is created and destroyed by our actions and the ways in which they get packaged together. As we will see, any such map of the terrain will not be very reliable for very long.

III. Ideal Theory Gets Overtaken by Events

There are compelling reasons from human history to view culture and society, especially the institutions and practices to be evaluated for their justice or lack of it, in this way. They have the character of reflexively co-evolving (packages of) strategies that constitute a changing topography of very temporary local equilibria among neighboring (packages of) strategies. To the degree that the evidence for this view of human political institutions especially is right, there is a very strong argument after all against the very idea of ideal theory and its alleged distinction from non-ideal theory in political philosophy. Schmidtz and others who reject the Rawlsian project built in part on this distinction will have a good argument against it. Certainly it gives them more than a gesture toward one.

It's not just that the reflexive interactions of strategies (and packages of coordinated strategies played in and by groups) continually shift the terrain of justice and present a moving target to ideal theory. The knowledge of these facts has to be built into the original position. There it makes ideal theory completely infeasible.

To see why, we need to recall what Rawls tells us about the veil of ignorance. Ideal theory begins with a deliberative process behind that veil of ignorance. But it is not really a veil against much ignorance. As Rawls says, behind the veil,

> [I]t is taken for granted . . . that [the parties] know the general facts about human society . . . . Indeed the parties are presumed to know whatever general facts affect the choice of the principles of justice . . . they understand political affairs and the principles of economic theory [not much help there]; they know the basis of social organization and the laws of human psychology [so they know more than we do]. Indeed the parties are presumed to know whatever general facts affect the choice of the principles of justice. There are no limitations on general information, that is, on general laws and theories, since conceptions of social cooperation must be adjusted to the characteristics of the systems of social cooperation which they are to regulate, and there is no reason to rule out these facts.[20]

[20] Rawls, *Theory of Justice*, 138. As Rawls reminds us in the first section of chapter IV, "Equal Liberty," section 31, "The four stage sequence," in the original position each agent is assumed to have enough knowledge to decide not only on the two principles of justice, but also on the characteristics of a just constitution, and similarly the features of just legislation.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

It's safe to say that if parties to the original position are omniscient about all material facts, it would be quite surprising were they to bargain to the arrangements characterizing the ideal theory Rawls argued for in 1971. Even the accretions to knowledge made since 1971 can be expected to make a difference to the bargain that parties make in the original position behind the veil.

Consider an obvious example of a change in our knowledge that has overtaken Rawls's claims about the difference between the natural and the social lottery in our deliberations behind the veil of ignorance. He famously tells us that "The natural distribution of [talents] is neither just nor unjust." And further, " . . . the distribution of natural assets is a fact of nature and . . . no attempt is made to change it, or even to take it into account."[21]

Consider what is now known about the nature of natural assets, and even more to the point, about natural disabilities and natural enhancements. Ideal theory tells us that reasoning leads to fair equality of opportunity through intervention in the social lottery. Now, however, ideal theory should surely provide equal reason to equalize regarding the natural lottery. Is this a mere wrinkle in ideal theory? If parties to the bargain factor in their knowledge of the social lottery and how to mitigate its impact, will they not likewise do so for the natural one once they acquire similar knowledge about how it works? Moreover, parties to the original position will have to reckon with the prospect of compensation not only for disabilities but also for enhancements that undermine fair terms of competition.

Think about the problem facing many professional baseball players in the period of widespread steroid use, or *Tour de France* riders during roughly the same period. This is an example of the way in which the playing field changes over time in ways that either overtake Rawls's original ideal theory or force it to change continually to accommodate updated information available in the original position.

Even if we start with Rawls's 1971 package of principles of justice, why should we think that after enough updates to our knowledge of the general facts — "the basis of social organization and the laws of psychology," — the resultant ideal theory of justice will bear much of a resemblance to the original one?

When it comes to constitutional and legislative strictures on commercial relations especially, the foresight demanded of agents in the original position will be weighty indeed. They will have to identify constitutional and legislative regimes that never provide perverse incentives to "game the system," no matter how social and technological relationships change. For example, they will identify arrangements that guarantee Nash equilibrium strategies among parties that preserve the difference principle. Failing to do so will defeat Rawls's objective of designing a sufficiently stable conception of justice, one that motivates the actions of agents. I owe this observation to Wayne Norman.

[21] Ibid., 101, 107.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

Another way to illustrate the manner in which the shifting landscape of temporary and local equilibria shape and reshape the conception of justice is to be found in the domain where *A Theory of Justice* seemed to so many people to pinch badly as Rawls tried to accommodate it to reality. One reason Rawls wrote *The Law of the Peoples* was presumably to argue that the ideal theory of justice as formulated for liberal democratic states could not be imposed as a normative obligation on other cultures and societies.

The Law of Peoples adds a compartment to ideal theory about how we, as individuals, and as liberal democratic peoples, are justly to deal with societies not so characterized. Some of these are decent societies, according to Rawls, even though, as in the case of theocracies for example, their institutions fail to comply with the conception of justice constructed in *A Theory of Justice*. Many readers of both of Rawls's books have been unconvinced if not offended by the free pass Rawls gives to such peoples that do not honor the package of principles articulated in *A Theory of Justice*. Ideal theory's toleration of other peoples' departures from Rawlsian strictures on justice is both a keynote of *The Law of the Peoples* and a point of deep contention. For example, Rawlsian ideal theorists who adopt a cosmopolitan conception of justice will perhaps accept Rawls's toleration as required by considerations of prudence. But they can hardly endorse his tolerance of merely decent societies as a moral requirement of ideal theory. Still another source of disappointment with *A Theory of Justice*'s noncosmopolitan approach, especially among Rawlsian egalitarians, is its silence on the transnational application of the difference principle. Aside from these objections, there are more fundamental ones that reflect the shifting landscape of social institutions across cultures and societies. If Rawlsian ideal theory, including its claims about other *peoples,* cannot accommodate these facts about the evolution of human civilizations, we may have another reason to find ideal theory otiose.

The moral agents in *The Law of the Peoples* are "peoples." Rawls derives several principles of justice among peoples from some sort of original position in which peoples find themselves. Now, it seems certainly to be the case that at some points during human history, including recent history, people have adopted strategies of various kinds that result in the coalescence of these strategies into packages that characterize national groups — ethnic, religious, even quasi-racial — and that enable their members to demarcate themselves as peoples, and to demarcate others as peoples different from themselves. Whether or not this is a morally regrettable pattern in human history, it is certainly not a permanent one. In the era of globalization, characterized by the abolition of barriers between individuals and the increased identification across these boundaries between individuals, identifications that trump their differences as "peoples," the claims of a law of the peoples to the status of ideal theory surely must be weakening, if it ever had any claims to begin with. In fact, one view of the history of the last two centuries is that the burgeoning of well-ordered constitutional democracies has been one of the causes of a weakening in the boundaries between peoples, one that has accelerated in recent years. Technology alone — the Internet's destruction of barriers to communication — homogenizes the moral norms of individual members of these democratic "peoples" and those of members of decent but decidedly nonliberal patriarchal hierarchical theocracies. In *some cases it goes further and dilutes the bonds that bind individuals together into peoples*, bonds that are arguably among the most injustice-producing norms in human history. One of the great accusations against "globalization" is its tendency to homogenize cultures. By and large, it has done so by making other peoples' cultures more like the one Rawls contemplated in *A Theory of Justice*. It is at least arguable that this tendency imposes the ideal and nonideal theory of that work universally, *pace The Law of the Peoples*.

It is reasonable to write into the original position knowledge of the impact of economic development, technological change, environmental degradation, climate change, and other barriers to national and cultural differences between people and peoples that do not violate the objectivity considerations to which the veil of ignorance caters.[22] Add to this the recognition that such changes provide us and other peoples with new and ever more egregious opportunities to depart from justice. Further, social science and the laws of human psychology may even lead us to believe that the forces that make for the organization of individuals into cohesive *peoples* are harmful to them, and to the existence of just institutions among them. When we do this, there seems less and less reason to take toleration of *The Law of Peoples*' additions to ideal theory seriously. Indeed, there seems little reason to take the ideal theory of *A Theory of Justice* seriously.

One might seek to defend ideal theory in the face of the moving landscape of injustices and their contraries by some scheme of conditionalization or qualification. Thus, ideal theory has a law of the peoples that is enforced if and when there are distinct decent nonliberal societies, but not otherwise: so, in a homogeneous world ideal, the law of the peoples drops out of the conception of justice or has no applicability. We could of course adopt a similar scheme for Rawls's *provisos* regarding complete compliance and favorable circumstances: the package of provisions that constitutes Rawlsian justice obtains just when there is complete compliance, favorable conditions, and cultural/ethnic/linguistic/racial homogeneity *cum* territorial integrity.

Suppose we ask the parties to the original position to include provisos for all these conditions in their calculations regarding the principles of justice. There seems no more reason to exclude historical, cultural, social considerations from the original position than the culturally local ones that, behind the veil, guide parties in bargaining to the Rawlsian conception.

[22] Cf. note 20.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

This understanding of the ideal theory of justice makes it a seriously unwieldy compendium of considerations contingent on the state of knowledge behind the veil of ignorance and the historical, cultural, social, and technological developments in which the parties can expect to find themselves. To accommodate all these considerations without explicitly mentioning them, ideal theory will have to be riddled with qualifications, *ceteris paribus* clauses, and frank admissions of indeterminacy of application.

IV. Can Rawls's Mature View Really Accord a Role to Ideal Theory at All?

Implicit in much of what I have written here is the claim that political philosophy is a completely inapt domain for anything like ideal theory. What is more, the trajectory of Rawls's own thinking in the years after writing *A Theory of Justice* reflects this fact — reflects it so well that we are left considering whether in the end Rawls would really have used the words to describe the historically, culturally, socially transitory recipe for justice he advocated in *A Theory of Justice.* To see why this might be so, let's consider ideal theories elsewhere.

Ideal theory seems aptly to characterize sets of claims in disciplines such as mathematics, and perhaps its applications in certain portions of physics, and perhaps even general equilibrium economics. Consider the ideal theory we all learn in high school — Euclidian geometry, or even better the Peano axioms in number theory we learn in mathematical logic. To begin with we have strong clear intuitions about many mathematical and some geometrical truths — so strong that in most cases we cannot even imagine their falsity. And the "we" of the previous sentence is pretty universal among educated persons over historical epochs actually interested in the subject. The firm agreement on mathematical truths leads us to search for a compendious body of theory that systematizes all or at least as many of these truths as we can contrive. (In arithmetic this task turns out to have limits undiscovered from the time of Archimedes to Gödel.) There are strong and widespread intuitions that mathematical and geometrical statements are absolutely true and known to be true. There is equally widespread agreement that mathematical and geometrical truths are not "about" physical facts, events, processes, entities, and so on, and that no empirical considerations could adequately substantiate or undermine them. So, we come to treat them as truths about a range of abstract, ideal objects, relations, and systems. In this sense mathematics is ideal theory.

In physics as well the term "ideal" also seems reserved for abstract and not concrete objects: we noted that the ideal gas law, PV = nrT is so called because it is strictly true only of gases composed of point masses between which no forces act, two things untrue of any concrete objects. Ideal theory

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

is broadly important in physics as a first step and as a heuristic device. As a heuristic device it has some computational or predictive usefulness and can be incorporated in measuring instruments or relied up when increasingly precise prediction is not required. As a first step, ideal theory identifies the ways in which theory is to be improved by reducing the levels of idealization. Thus, the history of the kinetic theory of gases is a history of successive reductions in the role of idealizing assumptions. A third use of ideal theory in physics is in the identification of limits to which nonideal physical processes or systems can attain but not transcend. A simple example is the ideal pulley in mechanics, a massless, perfectly rigid, and frictionless device for exploiting mechanical advantage. (Its properties are probably a logically inconsistent combination).

Beyond natural science, in economics, the proof of the existence of stable, allocatively efficient equilibrium in a perfect market is sometimes offered as another example of ideal theory. In this case too there is an intuition, or at least an intuitive argument, due to Adam Smith, that there is such an object — the perfect market. Its existence as an abstract object was only established after a century and a half of labor by mathematical economists. In all three of the cases of mathematics, physics, and economics, we have a good understanding of why actual concrete matters do not instantiate the truths of ideal theory. In the case of physics, and perhaps also economics, ideal theory only exists as a useful fiction — a pedagogical or a calculating device, which describes an unattainable state of affairs to which we may nevertheless aspire, since getting closer to it meets our aims.

By analogy with mathematics, one potential source for an ideal theory in political philosophy and ethics more generally would be a strongly shared set of intuitions about justice that admitted of a set of systemizing principles. But this is a conception Rawls rejects at the outset of *A Theory of Justice.*[23] There is no "irreducible family of first principles" of justice that are either self evident or detectable by a widely shared moral sense. The wide reflective equilibrium that is, according to Rawls, the result of deliberation in the original position is evidently not a matter on which all parties to the search for justice are in agreement.

Even if there were agreement on the principles of justice, the starting point of Rawls's derivation of the principles is strikingly different from the noncontroversial starting points of ideal theories elsewhere. In *A Theory of Justice*, Rawls adopts what he calls a "constructivist" method of which Samuel Freeman gives a cogent account:

> Kantian *constructivism* begins from a conception of the person and of practical reason, the ideal of free and equal moral persons who are both reasonable and rational. It "represents" or "models" this

[23] Ibid., 34ff. Cf. especially the long footnote, number 18 on that page.

> conception in a "procedure of construction," [in Kant, the categorical imperative, and in Rawls, the original position] . . . The principles chosen by the parties are objective so long as all who employ it reach the same or similar conclusions and the procedure incorporates all relevant requirements of practical reason. In this regard, moral principles are said to be "constructed" from a conception of the person and of practical reason.[24]

It is evident that the Kantian assumption that we are free agents endowed with powers of reason, moral autonomy, and objectivity, and are committed to self-realization is not one widely enough shared to confidently ground an ideal theory. To some it may sound more like a pious hope or a bit of sermonizing, rather like Rawls's claim that the "Aristotelian principle" is a psychological law:

> Other things equal, human beings enjoy the exercise of their realized capacities (their innate or trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity . . . We need not explain here why the Aristotelian Principle is true.[25]

Rawls himself came to realize that the method of Kantian constructivism was an inadequate one for a theory of justice by the time he wrote *Political Liberalism* (1993). By that time, as the title of an important paper by Rawls made clear, he had recognized that the enterprise of political philosophy was political and not metaphysical. In that paper he wrote,

> [J]ustice as fairness is intended as a political conception of justice . . . worked out for a specific kind of subject, namely for political, social and economic institutions . . . of a modern constitutional democracy. . . . Whether justice as fairness can be extended to a general political conception for different kinds of societies existing under different historical and social conditions . . . are altogether separate questions.[26]

Once the project of a theory of justice is treated as the problem of finding a "shared point of view among citizens with opposing religious, philosophical, and moral convictions, as well as diverse conceptions of the good,"[27] the search for ideal theory becomes otiose. Why?

[24] Samuel Freeman, *The Cambridge Companion to Rawls* (Cambridge: Cambridge University Press, 2003), 27. Brackets in the quoted material are Freeman's.

[25] Rawls, *Theory of Justice*, 426–27.

[26] John Rawls, "Justice as Fairness: Political Not Metaphysical," *Philosophy and Public Affairs* 14, no. 3 (1985): 223–51.

[27] John Rawls, *Collected Papers*, ed. Samuel Freeman (Cambridge, MA: Harvard University Press, 1999), 329.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

Because, as Rawls recognizes in *Political Liberalism,* there is no set of foundational normative claims on which all will concur. Instead there is a multiplicity of (comprehensive) conceptions of the good that each will seek to act upon. The members of that set do not share any common elements, or at least not enough common elements to construct a shared foundation for the conception of justice all parties will nevertheless agree on. The concept of justice will apparently not be based on some one single consideration or set of them that is embraced as a component of every comprehensive conception of the good. It will suffice if each such comprehensive conception of the good sustains the Rawlsian package for a *different* reason. These reasons might even be incompatible with one another, and so incapable of conjunction to form a coherent foundation for justice that ideal theory is supposed to provide. Moreover, as illustrated above, the cultures in which political choices are made and which give content to justice, are not written in stone, or these days even in indelible ink.

If justice as fairness is political, and not metaphysical, a matter of negotiation, after the veil is lifted, among people who know their conception of the good, what room is really left for ideal theory at all? Not much as far as John Rawls is concerned.

V. Is There any Role for Ideal Theory in Political Philosophy?

The previous section identifies three domains in which there is a role for ideal theory: mathematics, physics, and perhaps economics. Interestingly, ideal theories in all three domains share a common role in guiding "engineering" to desired, agreed-upon, precisely specifiable outcomes. If we want computers to perform correct calculations, we need to identify the programs — the assumptions of Peano arithmetic or Kolmogorov probability, for example, that they need to implement. If we want to move pianos vertically we need to know how more nearly to approach the ideal pulley and chain. If we want to produce the largest quantity of what people really want with all the available inputs or factors of production, we need to know the conditions under which the market functions to produce this result. Of course, in these three domains the precisely specified outcomes or objectives are unattainable. No hardware we can design will ever be free from breakdowns that result in outputs at variance with the mathematically right answer — the one given by its ideal theory. Physical theory — thermodynamics, material science, solid state physics — gives us the best reason to conclude that the ideal pulley is an unattainable object to which we can at most asymptotically approach.

The case is also the same in the theory of the perfectly competitive market. Though we have a proof of its allocative efficiency, that proof rests upon a half dozen assumptions that we know to be unrealizable in real markets — among them infinite divisibility, infinite numbers of buyers and sellers, complete futures markets, perfect information, infinite divisibility of commodities, constant returns to scale.

So, ideal theories in these domains share the features of "perfection," physical impossibility of realization, and standard-precisification. There is something else that these theories share in common: they identify a state of affairs that is impervious to human manipulation. Those who employ these theories may not want the mathematically right answer to a computation or a pulley that loses no mechanical advantage or a market clearing equilibrium. But thanks to ideal theory they know what these states of affairs would consist in, and there is nothing humans can do to prevent their attainment once the ideal assumptions on which they rest are satisfied. In all three cases, ideal theory identifies an objective fact about (perhaps abstract) reality whose existence is entirely independent of us and our aims, attitudes, or ambitions.

This much is obvious in the case of ideal theories in mathematics and physics. But it is also the case for the theory of the perfectly competitive market. The perfectly competitive market is proof to gamesmanship, to strategic manipulation, to attempts to corner it, destroy its informational efficiency, or otherwise to entrepreneurially exploit it for "rents." That is the whole point of the oft-expressed observation that in such a market everyone is a price-taker and no one is a price setter. The perfect market always attains an allocatively efficient outcome, no matter what attempts traders make to subvert it.

The actual markets that arise among humans are cases of what Friedrich von Hayek called "spontaneous order." They emerge repeatedly and independently in human history, and do so not only without human intervention or design, but in spite of it. It was Hayek's striking observation that free markets solve institution-design problems that humans do not recognize, cannot solve for themselves, and all too often seek to subvert. Of course actual markets don't do these things completely. They are not the perfect markets of ideal theory.

But justice and the foundations on which it rests are not much like any of these three domains. In fact, justice is so different from mathematical truths, mechanical systems, and spontaneously emerging social institutions, as to suggest that there is little scope for an ideal theory of justice that looks anything like ideal theory in these three domains.

To begin with, there do not seem to be widely shared norms of justice that admit of some sort of systematization by a pure theory. If there were, it would be likely that social and political philosophy would show the cumulation in its history from Plato to Rawls that mathematics and physics have shown, as mathematicians and physicists sought to unify and systematize the agreement on fixed truths in their domains.

As Rawls noticed late in his work, even when there is agreement on the justice of some arrangement, its grounds are usually controverted among those who agree about the justice of the arrangement. So, the search for

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press

axioms that systematize and unify, which makes good sense in pure theory, whether in mathematics or physics or even economics, has no scope even to get started in political philosophy.

Perhaps an even more obvious difference between justice and domains that admit of ideal theory is the imperviousness of these domains, even the domain of perfect competition, to human intervention.[28] Justice is the domain in which human intervention is both most crucial and most deforming. Martin Luther King, Jr. often said, and Barack Obama fondly quoted the observation that “the arc of the moral universe is long, but it bends towards justice.” If it is a factual claim, this view does not seem to be substantiated by history. If it were so substantiated, there might be scope for an ideal theory to explain at least how this is possible. There is no spontaneous order of justice.

As Rawls learned, through the evolution of his own thinking about justice from *A Theory of Justice* to *Political Liberalism*, what counts as a just political institution is very much a matter of culturally, socially, and historically changing norms, aims, and, most of all, strategies of individuals and groups. Each of us (individuals) and each of them (the groups to which we belong) are continually facing strategic interaction problems posed by the local institutions in which we and they operate. If justice (even perfect justice) ever obtains at any place and time, the outcome reflects temporary, transitory institutions that immediately begin to be changed by those who are affected and those who can take advantage of the outcome. If perfectly just political institutions were sufficiently like perfectly efficient markets, ideal pulleys, and the *abstracta* of mathematics, perhaps there would be a role for ideal theory in political philosophy. But they probably aren’t sufficiently like these ideal objects.

Philosophy, Duke University

[28] As Rawls himself recognized. Cf. John Rawls, *Theory of Justice*, 493. I owe this point to a referee.

https://doi.org/10.1017/S0265052516000376 Published online by Cambridge University Press