Utopophobia: On the Limits (If Any) of Political Philosophy, by David Estlund. Princeton, NJ: Princeton University Press, 2020. Pp. xviii + 379.

Over the past few decades, political philosophers have been entangled in a methodological debate over how theorizing about justice should proceed. Until recently, the dominant approach was ideal theory: theorizing about ideal or full social justice. But many have grown worried that ideal theory is practically irrelevant, telling us hardly anything about how to improve actual societies by meeting the various injustices we face. These self-described nonideal theorists have therefore called upon members of the discipline to shift their attention away from ideal theory, and toward nonideal theorizing about how to make existing societies more just. But ideal theorists resist this call, typically arguing that ideal theory enjoys a sort of methodological priority over nonideal theory. We need a conception of ideal justice, they claim, to serve either as a standard to approximate or as a long-term goal for reform.

In his innovative and wide-ranging *Utopophobia: On the Limits (If Any) of Political Philosophy*, David Estlund challenges not only many recent criticisms of ideal theory, but also the very terms of the debate. Much of the literature has focused on the alleged priority of ideal theory—on whether, in particular, it provides a standard or goal that we must appeal to when doing nonideal theory. But Estlund concedes that ideal theory enjoys no such priority, and that ideal justice plays neither role: it is a fallacy to assume that approximating it makes things better, and it is foolish and even dangerous to treat it as a goal if realistically we will never realize it. Nevertheless, Estlund argues that ideal theory identifies principles of ideal or full social justice that are moral requirements genuinely applying to actual societies, even if such requirements are unrealistic in the strong sense that human nature precludes their realization. And he argues that ideal theorizing about these requirements is important, in part because understanding them helps us to figure out how to improve actual societies, but also because this understanding is non-instrumentally valuable.

Estlund is a leading defender of ideal theory, so the publication of *Utopophobia* is a major event. Anyone interested in questions of methodology in political philosophy must reckon with it. Estlund's arguments are creative and persuasive, their conclusions are surprising and independently interesting, and the book is written in a crystal clear yet almost poetic style. One finishes each section knowing exactly what Estlund has argued, and usually without much to say in rebuttal. But while I find myself convinced of many of Estlund's particular theses, when I put together the pieces that comprise his central argument and consider its place in the broader dialectical landscape, it seems hard to believe that it will, or should, shift many allegiances in the ongoing ideal/nonideal theory debate. *Utopophobia* provides the strongest response to contemporary critics of ideal theory on offer. But if this is the best that can be said in favor of ideal theory, then, I think, the nonideal theorists have already won.

I.

Estlund begins, in Part I, with some major themes and distinctions. He is centrally concerned, he explains, with principles of justice, and these must be distinguished both from institutional blueprints and from practical proposals. First, whereas principles of justice are abstract moral requirements—say, that resources be equally distributed—a blueprint specifies the concrete institutions that would realize such principles. Second, whereas principles say what justice requires, a practical proposal advances 'a goal to set out for' (p. 10). These distinctions are important, since they allow Estlund to sidestep a significant objection to ideal theory that many critics (including this one) have recently mounted: that we shouldn't treat ideal justice as a long-term goal because this requires us to determine which institutions would be ideal and what steps would bring us toward them, but doing so is beyond

our epistemic capacity (Gaus 2016; Barrett 2020). Estlund agrees, but replies that we are nevertheless capable of identifying ideal *principles* of justice, and that such principles remain true even though we shouldn't treat their satisfaction as a goal. The aim of ideal theory is to identify abstract principles, not practical goals (ch. 1).

Part I also addresses other preliminary objections, including that, as 'political realists' argue, justice is not a moral standard (ch. 3), and that justice is essentially nonideal because the only agents who could achieve ideal justice would be moral exemplars to whom justice wouldn't apply (ch. 4). Estlund's discussion of political realism is clarificatory and persuasive, and his general point that even morally perfect agents would have conflicts of interests and disagreements, and so fall under the circumstances of justice, is well-taken. But granted that principles of justice are moral standards that would apply to morally perfect agents, actual humans are far from perfect, and it might seem that justice should therefore demand no more of us than we will ever achieve. Perhaps, in other words, principles of justice 'should bend to fit the shape of characteristic motives of human agents'—at least if these motives are given by human nature (p. 103).

Part II (chs. 5-7) provides an extensive argument against this 'bent view'. Although we should design practical proposals in light of what humans will do—lest they be 'utopian' in the pejorative sense (p. 11)—it is no objection to principles, claims Estlund, that we will never satisfy them. Humans may never meet certain purported standards of justice, say, because they are too selfish or partial. But this doesn't show that these are not genuine standards; it shows that humans are morally deficient.

As Estlund recognizes, the bent view may seem to follow from his position that principles of justice are requirements implying that we ought to do certain things, combined with the dictum that 'ought implies can'. His basic strategy is therefore to grant this dictum, but to argue, through a careful analysis of its application to a range of cases, that human motivation doesn't limit what we 'can' do—at least in the way proponents of the bent view contend. That I am too lazy to take out the trash, for example, doesn't show that I can't do it. And more generally, individuals can do things even when they are unwilling, or 'can't bring themselves', to do them, so long as they would tend to succeed if they tried. If humans are too selfish or partial to comply with principles of justice, this doesn't mean that humans can't comply with such requirements, but merely that they won't.

Estlund's discussion is subtle and illuminating, and his general point that theorists are often too quick to assume that motivational features block abilities (and so requirements) is hard to resist. But there is room to object to the details. For example, Estlund focuses on isolated cases, and it is unclear whether these generalize: even if, in any one-off case, I can act impartially in the face of temptation to selfishness, I may be unable always to do this, especially if I am confronted with temptation frequently and without respite. Many psychologists believe that willpower is a scarce though replenishable resource—that our capacity to exercise it is like a muscle that can fatigue and strengthen with training. If this is right, then the limits of human willpower would seem to constrain our abilities. Similarly, a long tradition of moral thought emphasizes the importance of habit to moral behavior. Perhaps being morally upright is like walking upright: even if someone with poor posture can, at any moment, snap themselves into an upright position, they may be unable to maintain consistent good posture without the appropriate habits (Dewey 1922). Plausibly, then, our habits, or the habits we can form, also constrain our abilities. But Estlund never considers such possibilities.

Estlund's arguments are nevertheless credible enough to place the onus on the defender of the bent view in developing this or some other reply. So let us set this worry aside, and turn to another, which

Estlund considers at length: that principles of full social justice require us to build and comply with certain institutions, but we shouldn't build such institutions given that we won't comply with them (even though we can). Such requirements seem therefore not to apply to us after all—not because we can't satisfy them, but because we shouldn't.

Part III addresses this concern by distinguishing between *concessive* requirements we ought to satisfy given that we won't do other things we ought to do, and *nonconcessive* requirements we ought to satisfy given that we will do everything else we ought. Simplifying somewhat, Estlund aims to show that both sorts of requirements can simultaneously apply despite appearing to conflict. To take a case that hits close to home: it can be both that Professor Procrastinate ought to accept an invitation to review a book and write the review by the deadline (a nonconcessive requirement) and that he ought not to accept the invitation given that he won't write the review in time (a concessive requirement). Or, more to the point, it can be both that we ought to build and comply with institutions (a nonconcessive requirement) and that we ought not to build those institutions given that we won't comply with them (a concessive requirement) (ch. 8). Estlund coins the term 'prime justice' (ch. 10) for nonconcessive requirements that society ought to satisfy given that it does everything else it ought (at least as far as justice is concerned), which take the form 'society ought to Build and Comply with institutions that meet the principles' of full social justice (p. 197). And he argues that actual societies are genuinely subject to such requirements.

Estlund's surprising argument that even though we shouldn't build institutions we won't comply with, we may nevertheless be under a requirement to build *and* comply with them, is convincing, deftly weaving through neighboring issues in deontic logic and the actualism/possibilism debate in ethics. And we can understand Part II as making the case that we can comply with those institutions that would, given full compliance, realize full social justice. Still, Estlund never considers an important objection: roughly, that principles of full social justice don't apply to us, not because we can't comply with or shouldn't build the institutions needed to satisfy them (call these 'ideal institutions'), but because we can't build such institutions.

My argument is simple. If we can build and comply with ideal institutions, then we can build them. And if we can build ideal institutions, then we must know how to build them, or at least be able to figure this out. But we don't know, and can't figure out, how to build ideal institutions. So we can't build them. And, by ought implies can, we are under no requirement to build and comply with them.

Estlund agrees that epistemic factors limit our abilities this way (p. 215). So the crucial premise is that we can't figure out how to build ideal institutions. But this, too, is something we have seen Estlund acknowledge, in Part I, when he dodges epistemic objections to ideal theory by suggesting that even though we can't identify ideal institutions or figure out how to achieve them, we can identify ideal principles (pp. 8-10). And although it remains open to Estlund to argue that we can indeed work out how to build such institutions, this would require him to engage with arguments to the contrary, rather than disregarding them as irrelevant. Even then, he would face a related worry. According to Estlund, requirements of justice apply to societies rather than to individuals, and so apply only if societies can meet them. But what can societies do? Estlund's discussion here, though helpful, never acknowledges that, in order for a society to do something, its members must be able to coordinate on it. Suppose you and I can each go to any number of locations. Does it follow that we can meet? Only if we can coordinate on some particular location. Likewise, a society can build ideal institutions only if its members can coordinate on which institutions to build (Stemplowska 2016). But this may be impossible, even for perfectly motivated individuals, given reasonable disagreements about principles

of justice and their institutional realizations. And this suggests, again, that societies can't build ideal institutions, and so are under no requirement to build and comply with them.

Putting things more carefully, this line of reasoning has two upshots. First, if we can't build the institutions needed to satisfy purported principles of full social justice because of our epistemic limitations or disagreements, then, by ought implies can, it follows that these are not genuine principles after all. So if my argument succeeds, it doesn't strictly speaking show that full social justice doesn't apply to us, but rather that full social justice must 'bend', if not to human motivational limitations, then to our epistemic and coordinative limitations. But second, recall that it is crucial to Estlund's defense of ideal theory that it aims only to uncover abstract principles of justice and not how to institutionally realize them. So it also follows that the sort of principles ideal theory delivers are unlikely to apply, because they are developed without consideration of such epistemic and coordinative limitations. Thus, while Estlund might rescue ideal theorists from the charge that they make unrealistic assumptions about human motivation, he leaves them susceptible to two other familiar charges of being unrealistic: that they ignore epistemic limitations and that they ignore persistent disagreements. And this time, the onus is on Estlund to show that such problems can be addressed.

II.

Even if Estlund can meet these challenges, another complication looms. I have noted that, according to Estlund, requirements of justice apply to societies rather than to individuals. This idea is somewhat obscure, so Estlund devotes Part IV (chs. 11-12) to motivating and explaining it. Estlund's basic thought is that requirements of social justice don't apply to individuals unconditionally, because though each of us ought to comply with them if others do, we shouldn't always comply when others don't. So if principles of justice are unconditional requirements, they must apply not to particular individuals, but to the group of individuals that constitute society—to society as such. But Estlund also compellingly argues that societies are not generally collective agents. So principles of justice don't seem to apply to any agent, individual or collective. This is a puzzling result.

As Estlund emphasizes, this is an instance of a general puzzle in moral theory, illustrated by his case of two doctors, Slice and Patch, who intuitively are required to save a patient's life: Slice should perform surgery on her and Patch should patch her up. The trouble is that neither will do his part (both are off golfing), but if either were to do his part without the other doing his, the patient would die more painfully. So each has a conditional obligation to do his part if the other does his, but since neither does his part, no conditional obligation is triggered, and neither violates an obligation. And yet, intuitively, the patient is wronged.

The puzzle is that the following claims are inconsistent: the patient is wronged, no agent has wronged her, but only agents are under requirements and so capable of wrongdoing. Estlund argues (convincingly) that Slice and Patch are not a group agent, and (plausibly but not decisively) that the patient is wronged though not individually by either Slice or Patch. So he concludes that we should reject the idea that only agents are under requirements or capable of wrongdoing. Slice and Patch are under a 'plural requirement' that applies to them not as agents, but as a group. And, Estlund claims, requirements of social justice are the same. Society is under a plural requirement to satisfy principles of justice, though none of the agents that constitute society are.

By itself, this doesn't resolve the above obscurity: we still need an analysis of plural requirements. Estlund proposes a conjunctive account. Slice and Patch are under a plural requirement just in case

both (i) each has a conditional obligation to do his part of the surgery if the other does his and (ii) it ought to be the case that—equivalently, it would be good that—both do their parts. Likewise, society ought to build and comply with certain institutions just in case both (i) each person in society ought to do so if everyone else does and (ii) it ought to be the case that everyone does this.

Estlund's treatment of this puzzle is fascinating, and perhaps accommodates our intuitions about Slice and Patch. But in the context of his broader argument, his solution disappoints. Estlund stresses the importance of full social justice being an unconditional requirement because he wants to explain its 'practical import'. Conditional obligations to comply with full social justice *if* everyone else does so are not enough, because such obligations are 'no more interesting or important than what we should do IF we could fly, or were physically impervious, or immortal' (p. 352, n.6, emphasis in original). Similarly, understanding social justice as something that ought to be, or would be good, is unsatisfactory for Estlund, since this renders it 'merely evaluative without any practical import' (p. 118). But if it is not interesting or important to say either that each agent has the relevant conditional obligation, or that it would be good if each complies, then why is combining these two elements and calling the conjunction a 'plural requirement'—as Estlund does—any better?

Estlund never addresses this question, and while there are certainly instances where two elements that lack value or significance on their own fare better in combination (ch. 14), this does not appear to be such a case. For, as Estlund argues, societies are not capable of having reasons (p. 220). So even if societies 'plural ought' to satisfy the requirements of full social justice, this is a peculiar sort of 'ought' that doesn't imply any reasons for action, and so lacks the 'practical import' Estlund seeks.

This problem might seem to generalize beyond full social justice, since, as Estlund notes, the puzzle of 'plural requirement' arises not only for ideal theorists, but for anyone who understands principles of justice (however ideal or nonideal) as unconditional requirements (p. 209). But the difference is that nonideal theorists needn't understand principles as unconditional requirements to explain their practical import. Instead, they may understand them either as non-requiring standards of comparative evaluation that rank societies from more to less just and so help us to set practical goals, or as nonideal conditional requirements whose satisfaction is a goal. Neither option is available in the case of full social justice, which, Estlund insists, is a requirement but not a goal. We therefore appear left with the conclusion that full social justice, in particular, is not especially interesting or important after all.

III.

Let us take stock. In defending ideal theory, Estlund disavows the mainstream position that ideal theory provides us with a practical goal or is in some other way prior to nonideal theory. Yet he contends that ideal theory identifies abstract moral requirements on societies. Specifically, Estlund argues—largely successfully—that we may be under requirements to satisfy principles by building and complying with requisite institutions, even if human nature ensures that we will never do this, and even if we shouldn't build these institutions given that we won't comply with them. But he faces the worry that the principles ideal theorists deliver don't apply because we can't build the relevant institutions due to our epistemic limitations or disagreements. And even if such requirements do apply, this just means that it would be a good thing if everyone satisfied them, and that individuals would have an obligation to do so if (counterfactually) everyone else did.

Suffice it to say, Estlund's defense of ideal theory, though chock-full of thought-provoking and persuasive arguments, doesn't leave the ideal theorist with much ground successfully unconceded.

Though he may be right that there are true principles of full social justice, ideal theorizing about such principles doesn't appear especially interesting or important given the paltry role its conclusions play in our normative economy. But this may seem unfair, since Estlund devotes Part V to demonstrating the value of ideal theory. So let us consider his major arguments there.

First, Estlund agrees with nonideal theorists such as Sen (2009) that comparative evaluations of justice—of 'more just'—are all we need to set practical goals. But he argues that without ideal theory we may be too quick to dismiss important achievements as unrealistic, that judgments about ideal justice sometimes inform comparative theorizing, and that 'pure comparativists' can't make sense of non-comparative judgments of 'unjust' or 'just' (ch. 13). These arguments, however, seem largely beside the point. Nonideal theorists don't swear off appeals to non-comparative judgments or thinking about unrealistic possibilities; they claim that working out how to improve justice should be the central *aim* of our theorizing, and deny that ideal theory 'tell[s] us much', or is 'needed, or particularly helpful' in this regard (Sen 2009, pp. 100, 102). And there is no risk of even thoroughgoing comparativists being unable to make sense of non-comparative judgments or intuitions, since these may be interpreted as implicitly referring to a comparison class—with 'unjust' meaning less just than some conceivable, feasible, or otherwise salient alternative, and 'just' meaning not unjust. Further, Estlund's account of how ideal theory facilitates comparative theorizing is that, precisely because many ideal theorists are wrong that approximating ideal justice always makes things more just (ch. 14), we can theorize about comparative justice by identifying 'countervailing deviations'—ways in which deviating from ideal justice in two respects makes things more just than deviating in one (ch. 15). But since deviations only sometimes countervail, judgments about whether one deviation countervails another by making a society more just presuppose comparative standards of justice. So there seems little practical benefit to identifying an ideal along with countervailing deviations instead of appealing directly to comparative standards.

To be clear, Estlund is surely right that ideal theory can sometimes facilitate nonideal, comparative theorizing. But this is unsurprising: so too can theorizing about dystopia, or science fiction, or almost anything that focuses our mind on social phenomena. And Estlund provides little reason to think that ideal theory is especially relevant in this regard.

This brings us to Estlund's most distinctive claim: that ideal theory is not valuable only in virtue of its practical relevance anyway (ch. 16). Instead, Estlund argues that understanding full social justice is non-instrumentally valuable, because this is part of being a virtuous person: good people are emotionally attuned to justice, and ideal theory contributes to this 'informed concern' (ch. 17). It is, however, far from obvious that virtue depends at all on understanding abstract principles of full social justice. And even if it does, Estlund appears to concede that someone who understands other (nonideal) truths about justice (say, about how injustice manifests and how to ameliorate it) but merely fails to understand full social justice is deficient in only a very narrow sense: they can't appreciate how far we fall from full social justice, and so can't have 'an informed concern for—a tendency to lament—the degree of injustice' (p. 326). Granting that lamenting injustice *to the right degree* is somewhat valuable, however, it doesn't plausibly have much value, let alone the 'great value' Estlund is after (p. 317). And while Estlund sometimes suggests that it is an understanding of his claim that full social justice may be unrealistic, rather than an understanding of particular principles of full social justice, which contributes to virtue (p. 324), this, even if true, doesn't bear on the value of ideal theory, which aims to discover the content of such principles.

Part V, then, fails to reverse our above conclusions. Ideal theory may somewhat facilitate nonideal theory, and understanding full social justice may have some limited non-instrumental value. But so what? If the central fault line of the ideal/nonideal debate were whether ideal theory has any value, then this would be a momentous conclusion. Yet this is not, I think, the best way to understand such a methodological debate. Perhaps there are extreme cases, but nonideal theorists don't generally claim that ideal theory is completely worthless or that no one should ever do it. Instead, they make an appeal—a 'practical proposal', in Estlund's terms—to political philosophers, beseeching the discipline to shift its focus away from ideal theory, toward the more interesting, important, and tractable questions of nonideal theory. Ideal theorists who insist that ideal theory enjoys methodological priority attempt a powerful reply: nonideal theory presupposes ideal theory. But Estlund, to his credit, recognizes that ideal theory enjoys no such priority, and so retreats to the weaker claim that ideal theory is merely of *some* value. This is hardly persuasive in the face of nonideal theorists' call for methodological change. Disciplines shouldn't shift their primary focus only when their old focus is worthless; they should shift when a new focus is on-balance better. So if the best that can be said in favor of ideal theory is that it tries to uncover principles that we shouldn't treat as goals and that don't imply any reasons for action, but that may occasionally facilitate nonideal theory and enable us to lament injustice to the right degree, then, again: the nonideal theorists have already won.

These criticisms notwithstanding, *Utopophobia* is an important book, full of intriguing ideas and arguments—some of which, we have seen, represent significant achievements, and many others I have been unable to discuss here. It provides the most careful and extensive defense of ideal theory to date. And though I have suggested that it won't, and shouldn't, change many minds about the broader ideal/nonideal theory debate, it remains essential reading to anyone interested in such topics.*

References

Barrett, Jacob 2020, 'Social Reform in a Complex World' in *Journal of Ethics and Social Philosophy* 17
Dewey, John 1922, *Human Nature and Conduct* (New York: Henry Holt and Company)
Gaus, Gerald 2016, *The Tyranny of the Ideal* (Princeton: Princeton University Press)
Sen, Amartya 2009, *The Idea of Justice* (Cambridge: Harvard University Press)
Stemplowska, Zofia 2016, 'Feasibility: Individual and Collective' in *Social Philosophy & Policy* 33

JACOB BARRETT
University of Oxford
jacob.barrett@philosophy.ox.ac.uk

* Thanks to Jerry Gaus, Christopher Howard, Sarah Raskoff, Greg Robson, and especially David Estlund for helpful feedback.