Hanno Sauer, *Debunking Arguments in Ethics* (Cambridge, Cambridge University Press, 2018), pp. xi + 244.

Published in *Utilitas*. 'Accepted Manuscript' version.

During the last two decades, a lot has been discovered experimentally about why, how, and when people make moral judgments such as 'abortion is morally wrong,' 'one ought to save five persons rather than one,' 'one ought to take care of one's children,' and so on. The moral philosophical question of the day is when and how these findings ought to change our substantive moral views or our views about the status of morality. Should we, as some debunkers demand, give up characteristically deontological moral views, or, as other debunkers would have it, merely switch our metaethical stance? Alternatively, should we be staunch anti-debunkers and remain unfazed by novel experimental findings about morality altogether?

Hanno Sauer addresses these questions in an immensely rewarding, stimulating, and crystal clear book that aspires to offer the first systematic discussion of debunking arguments in ethics. The book provides a conceptual map of different debunking arguments in ethics and provocative discussions (and mostly rebuttals) of existing debunking arguments in the literature.

Sauer's central thesis is that, though empirical findings can debunk or vindicate normative beliefs in principle, the conclusions of existing debunking arguments are frequently overblown. His meticulous discussion of extant debunking arguments provides a strong defence for his thesis, breaching new

territory with novel and creative solutions. In particular, his reflections about taking debunking arguments to metaethical, rather than normative, conclusions and the status of trolleyology are important additions to the debunking debate.

Philosophers interested in the prospects of these specific debunking arguments will have most to gain from Sauer's book. Before raising a critical point about the book's overall argument, I briefly summarise what I take to be each chapter's main points, along with some evaluative remarks. Given limited space, many nuances and arguments cannot, unfortunately, be discussed to the degree deserved.

The introduction nicely lays out the challenge posed by debunking arguments with the analogy of a gap between empirical findings and their alleged philosophical, normative consequences. Extreme reactions to the gap (to wit, claiming that *all* empirical findings ought to change our views or *none*) are patently absurd. So, the task is to explain *when*, exactly, an empirical claim about the origins of our moral beliefs jeopardises their epistemic status. Sauer's answer comes in three parts: a classification of debunking arguments, rebuttals of debunking arguments based on disagreement, and rebuttals of two debunking arguments against deontology.

Sauer begins by introducing five different genuine debunking schemes. Most schemes start with the premise that some belief that p is based on a process P and concludes that the belief that p is unjustified. They differ in the middle part, where we find premises about P failing to track truth (what Sauer calls 'off track' debunking), P being adapt to produce truth only in now irrelevant scenarios (obsoleteness), P leading to fundamental disagreement with epistemic peers

(symmetry), and P producing unacceptably large amounts of false positives or false negatives (detection error). Sauer explicitly aims to solely "illuminate the structure" (p. 35) of these schemes, and so he is not concerned with validity or soundness.

Throughout the book, Sauer sometimes refers back to these different kinds of debunking arguments. For example, he classifies Sharon Street's well-known debunking argument as being of the 'off track' type (A Darwinian dilemma for realist theories of value, 2006), and the classification should help readers unfamiliar with the debate to orientate themselves.

In the second chapter, Sauer turns to criticise Street's argument. According to Sauer's reading of Street, Street aims to debunk realism by showing that evolutionary explanations of morality suggest that our moral beliefs do not track mind-independent moral truths. According to Sauer, Street then confronts us with the choice to reject either of the three ingredients that lead to the dilemma: evolutionary theory, moral realism, or substantive moral judgments. Sauer's principal innovation in this chapter is to question Street's claim that the debunking argument can have solely *metaethical* rather than *substantive normative* implications. Street's success relies on presenting a set of *deeply held* normative judgments for *only* in that case, argues Sauer, would giving up moral realism indeed be the weakest link. Sauer thus concludes that Street's move to metaethical conclusions works only if the set of substantive moral beliefs that we would otherwise have to give up is sufficiently dear to us.

Sauer's discussion of the metaethical step is a creative and novel strategy to reply to some types of evolutionary debunking arguments. By placing the chapter

into the book's first, systematic part, he also suggests that his strategy about determining the target of debunking arguments (normative vs metaethical) generalises. Given the general lack of attention toward this issue, Sauer's contribution offers a fruitful starting point. However, important open questions remain. Most importantly, Sauer's strategy seems to apply only to selective debunking that threatens not all moral judgments. In that case, the targeted moral judgments might indeed seem easy to give up, compared to the cost of relinquishing moral realism. Since extant selective debunking arguments target moral judgments that seem easy to give up, compared to giving up realism, it is unclear where Sauer's strategy could apply.

Moreover, his reply to Street is ultimately not convincing because it does not address the strongest construal of Street's argument. It seems false to suggest that Street's debunking argument, and others like it, depend on pragmatic considerations about what theory to choose. A more charitable interpretation would interpret it as confronting realists with an epistemic problem: *if* realism implies that moral truths are so-and-so *and* that we have moral knowledge, then showing that evolution is incompatible with the knowledge-claim defeats realism, though for epistemic, not pragmatic reasons.

In the third chapter, Sauer introduces the distinction between depth and scope in debunking arguments. The *scope* of a debunking argument can be global (concerning all moral judgments) or selective (concerning only a subset of moral judgments) and further classifies the debunking types identified in chapter 1 along these two dimensions. The debunking argument's *depth* is determined by how strong its target moral judgments are affected, that is whether they are

"thoroughly," "frequently," or only "somewhat" unreliable (p. 75). Apart from classifying some existing debunking arguments, Sauer discusses the prospects of global, deep (thorough) debunking arguments and selective debunking arguments. About the former, he flirts with so-called pre-established harmony explanations, suggesting that there is little reason to believe that *all* moral judgments are thoroughly unjustified (p. 84). About the latter, he is even more sceptical, mainly because he argues that selecting debunking can easily overgeneralise.

Chapters 4 and 5 are concerned with moral disagreement broadly construed. In chapter four, Sauer discusses an instance of the well-known argument of disagreement against moral realism due to Doris and Plakias (How to argue about disagreement, 2008). Accordingly, there are some fundamental moral disagreements, that is, both parties to the disagreement are sufficiently well informed, make no cognitive errors, and so on, and yet they disagree. Given the evidential symmetry between both disputants, anti-realists claim, both should revoke their belief. Sauer argues that the argument does not succeed. Rather than rehearsing common objections, Sauer's argues that it is not enough for anti-realists to show that there is *some* fundamental moral disagreement, or even that *very many* of our moral beliefs are subject to fundamental moral disagreement. Rather, anti-realists have to show that there is fundamental disagreement about *normatively significant issues*, which he clusters around three major themes: moral egalitarianism, normative individualism, and opposition to gratuitous violence (p. 111). Leaning on recent optimist arguments about collective moral progress, Sauer then reviews some empirical findings to show that there is more and more agreement, rather than disagreement, about these issues.

Chapter 5 is based on an earlier paper of Sauer (Can't we all disagree more constructively?, 2015) and deals with political disagreement and the argument, which Sauer attributes to Haidt (The righteous mind, 2012), that liberals ought to take into account the intuitions of conservatives to resolve the disagreement. Sauer excels in this chapter. Through minute attention to the empirical findings, he shows that Haidt's argument succeeds only if both Social Intuitionism and Moral Foundations Theory are true. Sauer shows that both cannot be true at the same time and so the debunking of Liberalism fails.

Chapters 6 and 7 turn the focus to debunking arguments of deontology. Again, Sauer discusses particular instances of debunking arguments, and he argues that neither succeeds. In chapter 6, Sauer argues that the findings of trolleyology are not to be taken seriously and that, therefore, debunking arguments based on such experiments fail to bridge the gap between empirical findings and normative implications. He arrives at what is probably the book's most controversial conclusion as he finds that "it might be high time to retire [trolleyology] to where it belongs: … in the line of a runaway trolley, with no one there to stop it" (p. 174). Roughly, proponents trolleyology argue as follows. Experiments that posit sacrificial dilemmas show that subjects chose to save the greater number of people in some cases, but that they do not choose to save the greater number in other cases. However, there are no morally significant differences between cases where subjects save the greater number and where they do not. So, proponents of trolleyology-debunking claim, the latter judgments must be deficient. Since they are characteristically deontological moral judgments, trolleyology-debunker conclude that deontological moral judgments are

unjustified. For the argument to yield a sweeping conclusion about all deontological moral judgments, the experiment must, of course, be ecologically valid: subjects' judgments in trolley experiment must be good indicators of moral judgments in general. Sauer denies this. Based on an assessment of the normal functioning of moral intuition and judgment, he argues that trolley experiments disqualify themselves by introducing *novel*, overly *specific* scenarios that create *imaginative resistance* and whose outcomes are *certain* in a way that the outcomes of ordinary moral judgments rarely are (pp. 160-8). A popular reply to this charge is that the core finding of trolleyology lies in the *difference* between moral judgments in normatively identical situations; thus, ecological validity is not per se an issue. Sauer rebuts this reply by arguing that trolleyology' s lack of ecological validity also sharply diminishes the reliability of inferences from the findings about differential responses.

Sauer's rebuttal of trolleyology will jibe well with critics of such experimental moral philosophy, who have long suspected that findings based on contrived experimental settings do not incriminate 'real' moral life. Sauer's detailed discussion of the criteria for ecological validity in the moral case is the best defence of this position available. Proponents of trolleyology, however, will wonder whether Sauer is taking it a bit too far. First, though bizarre trolley cases (to which Sauer's criticism may apply) are often highlighted in discussion, supplementary materials reveal more quotidian cases that seem much closer to mundane moral reality. It is less clear that Sauer's criticism applies to these cases, too. Second, for Sauer's sweeping conclusion to hold up, it has to be the case that experimental settings allow no inferences about ordinary moral judgment whatsoever, which

seems overly strong. Perhaps Sauer thinks that scientific approaches other than, say, experimental psychology are better suited to study morality or at least necessary complements to it. Confining trolleyology altogether to the heap of failed approaches, however, may be too ardent a response.

In chapter 7, Sauer turns to a discussion of the side-effect effect, according to which people asymmetrically attribute various agential features (e.g. intentionality) to other agents when something normative is at stake (p. 178). The side-effect effect has been used in attempts to debunk characteristically deontological intuitions and, in an updated version of an earlier argument (It's the Knobe Effect, Stupid!, 2014), Sauer offers a unifying explanation of the effect aimed at defusing the debunking challenge.

The final chapter spells out the broader implications of the book's arguments and entertains the tantalising idea of vindicating arguments, which Sauer construes as being based on several failed attempts to debunk some set of moral beliefs. Many more arguments and nuances of Sauer's book merit more thorough discussion: they offer both tasking challenges for debunkers, and fruitful, rewarding pointers toward improving their arguments.

Despite the book's strength in criticising existing debunking arguments, it leaves unanswered some questions about the soundness of debunking schemes in general. That is because the book focuses on the current empirical record's philosophical implications, not on the more general question of how or when empirical findings can have normative (or metaethical) conclusions at all. So, though the title suggests a general, systematic discussion of debunking arguments, the book is more focused on criticising specific arguments (with

fruitful results), while important questions relevant to the nature of debunking remain unanswered. To begin with, Sauer's main innovation about the nature of debunking is his classification and the discussion of the depth/scope variables. While newcomers to the debate might find the classification helpful, Sauer fails to explain why his classification ought to be preferred over alternatives in the literature (e.g. the 'truth-debunking' vs 'justification-debunking' distinction). There are also omissions (for instance, there is no discussion of ontological, parsimony-based debunking arguments) and some schemes also appear to be mere instances of one another. For example, consider Sauer's off-track debunking: some influence X shaped belief-forming process P, and therefore P does not track the truth of some set of beliefs. That seems relevantly similar to Sauer's obsoleteness debunking: P is culturally or biologically adapted to produce judgments only in environment H* but not in H. In that case, P also seems to fail to track the truth of some set of beliefs. Though very different experimental sources may fuel both challenges, their structure is relevantly similar from an epistemological point of view. Much focus of the recent debunking debate is to understand more about this scheme. What, for example, does it mean for some process to be 'off track,' and what could be experimental evidence for that? Sauer avoids some of these questions by focusing on debunking arguments that target the justification of moral beliefs, not their truth. Though some have recently argued that *all* debunking arguments ultimately target justification, that itself is a substantial claim about the workings of debunking arguments that will have to be part of a full debunking theory. Ultimately, Sauer's book assumes that some connection

between the causal explanation of a belief and the belief's epistemic justification exists, though it does not elucidate what that connection is.

The book's strength is thus, as already mentioned, its discussion of the empirical premises of debunking arguments, assuming that debunking schemes are valid and sound. However, given that Sauer takes himself to have refuted or at least cast serious doubt on every single debunking argument he discusses, the reader may be left wondering when, exactly, "debunking arguments work" (p. 218).

Sauer's creativity, deep familiarity of the empirical material, and elegant, approachable style make this an exciting and important book on one of moral philosophy's important topics. It is also a joy to read. Researchers working on naturalistic approaches to morality will benefit from it immensely, and its approachable style and philosophical depth make it eminently useful for use in graduate and undergraduate seminars on the topic.

Michael Klenk, Delft University of Technology