Can Moral Anti-Realists Theorize?

Michael Zhao

(forthcoming in *The Australasian Journal of Philosophy*)

Abstract

Call "radical moral theorizing" the project of developing a moral theory that not only tries to conform to our existing moral intuitions, but also manifests various theoretical virtues: consistency, simplicity, explanatory depth, and so on. Many moral philosophers assume that radical moral theorizing does not require any particular metaethical commitments. In this paper, I argue against this assumption. The most natural justification for radical moral theorizing presupposes moral realism, broadly construed; in contrast, there may be no justification for radical moral theorizing if moral anti-realism is true.

Many moral philosophers think that moral inquiry should be guided not just by conformity to our existing moral judgments, but also by considerations like consistency, simplicity, unity, and explanatory power. In other words, they don't just want a codification of our existing moral judgments; they want a genuine moral *theory*, a view that exhibits various theoretical virtues to a high degree. I'll call *radical moral theorizing* the project of developing such a theory.

Philosophers who engage in radical moral theorizing typically do so in isolation from substantive metaethical inquiry.[1] The assumption seems to be that radical moral theorizing doesn't require any particular metaethical commitments:

I would like to thank Robert Audi, Harjit Bhogal, Camil Golub, Kris McDaniel, Sam Scheffler, Sharon Street, and two anonymous referees for invaluable comments on earlier drafts of this paper.

[1]As evidence for this claim, consider the lack of substantial engagement with metaethical questions in the following selection of prominent theory-heavy works in normative ethics in the past few decades, which I treat as representative. Derek Parfit (1984, 446) mentions metaethics only briefly, while discussing the question of whether a self-effacing theory could still be true. Shelly Kagan (1991, 262) similarly mentions metaethics only once, in discussing a principle that

regardless of whether you're a moral realist, a non-cognitivist, a fictionalist, or some other kind of anti-realist, you're justified in seeking a consistent, simple, unified moral theory. In this paper, I'll argue against this assumption. On my view, the most natural justification for radical moral theorizing requires moral realism, broadly construed, and there may be no justification for it if realism is false. This does not mean that realism is *necessarily* friendly to radical moral theorizing, however; as I'll show, whether a particular version of realism is will depend on things like how much it prizes parsimony and how much confidence it has in our pre-theoretic moral judgments.

Here's an outline of the paper. In §1, I'll say what I mean by "realism" and "radical moral theorizing" in more detail. In §2, I look at the question of what is supposed to justify the emphasis on theoretical virtues like simplicity and explanatory depth. Drawing on an analogous discussion in the philosophy of science, I argue that a natural justification for wanting our moral theories to have these features is that they are conducive to truth, understood as correspondence to the moral facts. This is precisely what moral realism (in my sense) accepts, so this shows that moral realists have a proprietary, though defeasible, justification for radical moral theorizing. Next, I look at the auxiliary assumptions that one must accept for the justification to work, and show that disagreement about these assumptions among realists of different stripes explains disagreement about radical moral theorizing. I close the section by arguing that quasi-realists, who use deflationism to legitimize talk about moral truth and fact in a non-cognitivist framework, cannot appeal to the same justification. Finally, in §3, I argue that several natural attempts that are available to the anti-realist to justify our concern for features like simplicity, depth, and even (to some degree) consistency fail. In the absence of further arguments, radical moral theorizing is likely unjustified if moral realism is false.

1 Preliminaries

To start, let me say what I mean by "moral realism" and "radical moral theorizing" in more detail. I'll use "moral realism" in a broader way than usual, to mean simply the thesis that there are moral truths, in some metaphysically demanding sense of "truth."[2] Realism, in my sense, includes not just what traditionally goes

morality must take the nature of persons into account. And Brad Hooker (2003, 14f) mentions non-cognitivism and error theory in passing, simply to state his neutrality on issues of metaethics.

[2]This usage accords with Geoffrey Sayre-McCord (1986)'s definition of "moral realism," the thesis that moral claims are truth-apt and often true. Mark van Roojen (2015, ch. 2) uses the term

by that label, the thesis that there are *mind-independent* moral truths, which hold independently of our attitudes, beliefs, and practices, but also various forms of *non-objectivism*, according to which there are moral truths, but they are ultimately mind-dependent. In contrast to what I have called moral realism are varieties of *anti-realism*, which deny that there are moral truths, in any inflationary sense of "truth." This will include various forms of *error theory*, according to which moral claims make false presuppositions, so are uniformly untrue, and *non-cognitivism*, according to which the function of moral claims is not to describe any moral facts, but to express the speaker's non-cognitive attitudes.

And by "radical moral theorizing," I have in mind an approach to normative ethics that places a heavy emphasis on features that are sometimes known as "theoretical virtues," like simplicity, generality, unity, and overall coherence. Radical moral theorizing begins with our existing moral judgments, those that we are disposed to accept before engaging in moral theorizing, and tries to develop a systematic moral theory by finding a set of basic principles that exhibit various theoretical virtues, from which we can derive a large portion of our pre-theoretic judgments. The theoretical virtues that these principles are taken to need to have vary,[3] but I'll focus on the trio of *consistency*, *simplicity*, and *explanatory depth*. A set of foundational principles is *consistent* if it generates no contradictions; it is *simple* if it is small in number and employs few basic moral concepts; and it is *explanatorily deep* if the foundational principles are distant enough from facts about specific cases that they provide explanatory understanding of those specific facts. Obviously, simplicity and depth come in degrees: one theory can be simpler or deeper than another theory, even if both are simple or deep to some degree.

Views like utilitarianism and Kantian deontology embody simplicity to a very high degree. They both recognize a single foundational principle: the principle of utility or the categorical imperative. T. M. Scanlon (1998)'s contractualism is simple, but less so than utilitarianism and Kantianism, since it purports

"minimal moral realism" in roughly the same way. The condition that there be moral truths *in a demanding sense* rules out metaethical views that merely recognize the existence of moral truths in a *deflationary* sense of "truth," the sense in which any declarative sentence "p" is equivalent to "it is true that p."

[3]Kagan (1991, ch. 1) lists simplicity, power, and overall coherence as features that we want moral theories to have; Hooker (2012) lists, among other things, generality, foundational unity, codifiability, and commensurability; Timmons (2013, ch. 1) lists consistency and explanatory power as standards for evaluating moral theories; Driver (2022, §2) lists systematicity, depth, simplicity, coherence, and accuracy as theoretic virtues.

only to be an account of the part of morality that concerns our interpersonal obligations. But within this part, it also posits a single principle, that an act is wrong if and only if it would be forbidden by any set of principles that no one could reasonably reject. All three theories are deep: in the case of each, the foundational principle is far enough removed from moral facts about specific cases that it helps us understand why particular acts are right or wrong. Finally, given that there is a single, straightforward foundational principle for each of these theories, we have good reason to think that they are consistent.

Views like these are sometimes taken to be the outputs of radical moral theorizing: even though these theories are revisionary—that is, they frequently conflict with our existing moral judgments—the premium that radical moral theorizing places on consistency, simplicity, and explanatory depth means that it favors these theories over ones that conform better to our existing moral judgments, but lack these theoretical virtues. That is to say, proponents of radical moral theorizing assume that these theoretical virtues frequently outweigh another virtue: *goodness of fit*, or conformity of a moral theory to our existing moral judgments. Radical moral theorizing thus contrasts with more conservative forms of moral inquiry, which take goodness of fit to be much more important. The deliverance of a more conservative approach might resemble a Ross-style pluralism that recognizes many distinct types of *prima facie* duties, which cannot be unified at any more foundational level, or a kind of particularism, which does not recognize any neat and absolute principles.

2 Realism and the theoretical virtues

The emphasis that radical moral theorizing places on features like consistency, simplicity, and depth might seem strange or even fetishistic, especially if it comes at the expense of goodness of fit. We might sympathize with W. D. Ross (1930, 23) when he wrote, "loyalty to the facts is worth more than a symmetrical architectonic or a hastily reached simplicity." Why should we care that a moral theory exhibits these features in the first place? Moral philosophers are typically reticent about this question; some of them simply find it obvious that we should.

In this section, I'll give a natural answer to these question: that these features are conducive to truth, in the sense of correspondence to facts within some domain. And this presupposes exactly what I've called moral realism. (At the end of the section, I'll argue that *quasi-realism*, a form of non-cognitivism that tries to mimic realism, can't give the same answer.) This doesn't automatically mean that anti-realists can't justify caring about these features; in the next section, I'll

look at defenses of the theoretical virtues that are available to anti-realists. Nor does it mean that realists necessarily have all-things-considered reason to engage in radical theorizing. As the argument will show, whether it makes sense to go in for radical theorizing also depends on other issues that different realists take different stands on. Nonetheless, the upshot of the argument is that realists have a proprietary, although defeasible, justification for engaging in radical moral theorizing.

2.1 The case of science

I'll start by looking at a parallel question in the philosophy of science: why prefer scientific theories that are consistent, simple, and explanatorily deep, particularly when they purchase these features at the cost of conflict with our observations?

First, note that there is an obvious justification for consistency: theories that are inconsistent cannot correspond to the facts, since reality itself contains no contradictions. So consistency is conducive to truth, understood as correspondence to reality, in being a necessary condition for it.

What about defenses of other theoretical virtues? Many philosophers are pessimistic that the status of simplicity and depth as theoretical virtues can be justified. In the case of simplicity, Steven French (2014, 57) writes, "it is more or less accepted that there is no argument that demonstrates that simplicity tracks the truth." But other philosophers are more optimistic. Here, I'll rehearse a justification by Elliott Sober (2015, ch. 2), based on an argument from Hans Reichenbach, for a certain kind of simplicity and explanatory depth, a preference for *common-cause* over *separate-cause* explanations of correlations: for example, favoring the explanation that a single force governs both the attraction of earthbound objects to Earth and the attraction of astronomical objects to each other over the explanation that there are two separate forces at work. Or, to take a mundane case, favoring the explanation that two students who turned in the exact same essay wrote it together over the explanation that they each happened to write the same essay independently. Common-cause explanations are simpler than separate-cause explanations, since they posit fewer explanantia; they are also deeper, in an important sense, since they reveal phenomena to be unified at a deeper level than separate-cause explanations do.

The basic idea is that *the likelihood of the evidence on the common-cause hypothesis is much higher than on the separate-cause hypothesis*. In other words, the probability that two students would turn in the exact same essay given their having worked together is much higher than the probability that they would turn in the exact same essay given that each worked independently, which is basically nil.

Using Bayesian reasoning, as long as we suppose that the separate-cause explanation is not *drastically* more plausible beforehand than the common-cause explanation, then the common-cause explanation is more probable than the separate-cause explanation given the evidence.

I want to note the realist assumptions that these justifications employ. We should favor consistent theories because *reality contains no contradictions*; we should favor common-cause explanations because *it is more likely that a single event would generate the correlation than two separate ones would*. These justifications assume some domain of facts, and proceed by showing that the theory is more likely to be true—to correspond to the facts—if it has the features. In contrast, it is entirely unclear how we would make sense of these justifications at all if we were anti-realists about the domain that science purports to describe: it is unclear how we would make sense of the idea that it is more likely that a single event would generate a correlation than two separate events would if we took the events to be theoretical posits, for example.

These justifications of the theoretical virtues do not yet show why we are sometimes permitted to sacrifice goodness of fit in their name. Why is it permissible to revise a particular observation if some consistent, simple, and deep theory renders it false? Well, there's always the possibility of error in our observations. Roughly, if we think that it is more likely that the particular observation is false than the theory conflicting with it, we should revise the conflicting observation to be consistent with the theory. More precisely, the degree to which we should side with theoretically virtuous theories over particular observations in a conflict depends on how much simplicity and depth increase the probability that a theory is true, on the one hand, and how much conflict with our observations decreases that probability, on the other. But unless it is extremely unlikely that our observations are false, there will be cases in which a virtuous theory that conflicts with some of our observations is more likely to be true than a less virtuous theory that has a better fit. Such a justification of the theoretical virtues allows us, under certain conditions, to ride roughshod over offending observations in their name.

2.2 The case of ethics

So much for the case of science. How would analogous justifications go in the case of ethics? Consider a similar Bayesian justification for simplicity and explanatory depth in the form of preferring a unitary explanation to a pluralistic one of a body of specific moral facts.

Take the fact that injuring someone is wrong, that breaking promises is wrong,

that failing to reciprocate kindness is wrong, and so on. One possible explanation for these moral facts, analogous to a separate-cause explanation, is that there are many kinds of moral duties, which cannot be unified at any deeper level; each fact is then explained by its being covered by some kind of moral duty. This would be the form of non-unitary explanation favored by a pluralist like Ross. Another explanation, analogous to the common-cause explanation, is that these acts all are wrong in virtue of their having some feature, and that feature has constant moral relevance; for example, perhaps these acts typically cause suffering, and the fact that an act causes suffering makes it wrong. Let's assume for now that some version of both hypotheses are equally consistent with the moral facts; I'll consider below how the argument fares if the unitary hypothesis conflicts with some of them. Clearly, the unitary hypothesis is simpler, and possesses more explanatory depth than the hypothesis that posits disparate kinds of duties. Obviously, if both hypotheses are equally consistent with the facts, many philosophers would prefer the unitary hypothesis in cases like this. But why?

The basic idea is that the pluralistic hypothesis is *compatible with a much broader range of possibilities* than the unitary one is. If the pluralistic hypothesis is true, then in the absence of knowledge about the specific moral facts, perhaps injuring others is wrong, but not reciprocating kindness is permissible, breaking promises is permissible, …; or perhaps injuring others is permissible, but not reciprocating kindness is wrong, but breaking promises is permissible, …; or any one of a large number of other possibilities. It would very unlikely, if all we knew were that there are many kinds of moral duties, that injuring others, not reciprocating kindness, breaking promises, and so on, would all be wrong. In general, if there are n possible moral duties, then there will be 2^n possible sets of moral facts, of which only one is consistent with the actual facts. On the other hand, the unitary hypothesis is compatible with a much smaller number of possibilities, so the likelihood of the actual moral facts given the unitary hypothesis is much larger: the only possibilities that are compatible with it are ones where, for a given feature, every act with that feature has the same moral status as every other act with that feature. So if there are m possible morally relevant features, then there will only be m possible sets of moral facts, of which one is consistent with the actual facts.

Now, Bayes' rule tells us to assign credences to hypotheses that are proportional to our initial credences in them and how likely the data are on each hypothesis. The fact that the actual moral facts (the data) are much likelier on the unitary hypothesis than on the pluralistic one means that, unless we take the pluralistic hypothesis to be *drastically* more plausible beforehand than the unitary

one, the unitary hypothesis (which is simpler and deeper) is more likely to be true.

Of course, this argument assumes that both hypotheses are fully consistent with the particular moral facts, and that we can take the facts as given. But we can relax both assumptions: we can assume instead that what we possess are not specific moral facts, but rather our moral *judgments* about specific cases, which are fallible; and that the unitary hypothesis conflicts with those judgments more than the non-unitary hypothesis does. If our judgments are fallible, then failure to account for all of them will not be a lethal flaw; rather, all we can say is that a hypothesis that accounts for more of them is, all else equal, more likely to be true than a hypothesis that accounts for fewer of them. But note the two hypotheses are not equal, since one hypothesis is much simpler and deeper than the other, and we might antecedently prefer a simpler or more complex hypothesis. So we have to make a trade-off between these different considerations. Which hypothesis we should prefer depends on three things: first, how conducive simplicity and depth are to truth; second, how much we antecedently prefer complex hypothesis to simple ones; and third, how confident we are in our pre-theoretic moral judgments. The first is simply a given, a function of how much more restricted the possibilities compatible with the simple hypothesis are than those compatible with the complex one; but the other two are things that we can have different views on. If we have a strong antecedent bias toward complex hypotheses, or are very confident in our pre-theoretic judgments, then the more conservative and complex hypothesis likely has a higher chance of being true than the more revisionary and simpler theory. On the other hand, if we do not have a strong antecedent bias toward complex hypotheses, or think that our pre-theoretic judgments have a decent chance of being false, then the more revisionary and simpler hypothesis may well have a higher chance of being true.

Again, note the realist assumptions behind this justification of valuing simplicity and depth: we should prefer simpler and deeper to more complex and shallower moral theories, even potentially at the cost of conflicting with more of our existing moral judgments, because *the simpler, deeper theory is likelier to be true* in virtue of *being more likely to correspond to the moral facts*. This justification assumes that there is some domain of moral facts, and that moral claims are true in virtue of corresponding to these facts. This is exactly what moral realism (in my sense) claims, so moral realists have a natural justification for theoretical virtues like simplicity and depth.

2.3 Pro- and anti-theory realists

Now, this is not to say that moral realism will automatically be friendly to radical moral theorizing. Rather, the argument merely shows how realists *might* be allowed to go in for such theorizing. And it allows anti-theory and pro-theory realists to locate the sources of their disagreement. As we mentioned, different realists might have different views on two separate issues that affect whether a simple, deep, and revisionary theory is more likely to be true than a complex, shallower, and conservative one: (1) how much they antecedently prefer a simple or complex theory, and (2) how likely it is that our pre-theoretic moral judgments are false.

And we find that two major fault lines among realists in terms of friendliness to radical theorizing lies where my justification of such theorizing would predict. First, consider the difference in emphasis that different realists place on ontological parsimony. Moral naturalists tend to have a strong commitment to parsimony, antecedently preferring simple theories to complex ones, since they believe that the natural world tends to admit of simple explanations, and that ethics is in the business of describing part of the natural world. In contrast, moral non-naturalists have a much weaker antecedent commitment to parsimony: since the moral domain is entirely separate from the natural world, there is no reason why we should expect the former to be simple. The justification for theorizing that I have offered correctly predicts that naturalists are friendlier to theorizing; in contrast, many non-naturalists tend to be moral particularists, who categorically reject any unified theory.

Second, consider the difference in confidence that different realists have in our pre-theoretic judgments. In one camp, consider classical intuitionists like Prichard and Ross, who had a great deal of confidence in them. As Ross (1930, 29f) wrote,

> That an act, *qua* [instance of a duty] is *prima facie* right is self-evident; ... in the sense that when we have reached sufficient mental maturity and have given sufficient attention to the proposition it is evident without any need of proof, or of evidence beyond itself. It is self-evident just as a mathematical axiom, or the validity of a form of inference, is evident.

According to the justification of theorizing that I have offered, it is no accident that Prichard and Ross delivered not a revisionary moral theory, but rather a conservative systematization of commonsense morality: given that it is very un-

likely that our judgments about various *prima facie* duties are false, a theory inconsistent with them is less likely to be true even if it is simpler or deeper. In contrast to these realists, consider a realist like Michael Huemer (2007), who accepts certain aspects of intuitionism while thinking that our intuitive moral judgments have a substantial chance of being false: they might be the products of parochial biases, framing effects, or other unreliable processes. Similarly, it is no surprise that Huemer advocates for a revisionary approach to normative ethics: given that our intuitive judgments might very well be false, a theory may not sacrifice too much plausibility by conflicting with even a large set of them.

The upshot of my argument is that although moral realism isn't *necessarily* friendly to radical moral theorizing, it provides a defeasible reason to engage in it. Roughly, a particular realist should go in for radical theorizing if and only if he antecedently thinks the moral truth is more likely to be simple than complex, or lacks high confidence in our ordinary moral judgments.

2.4 Can quasi-realism employ the same justification?

At this point, one might wonder whether *quasi-realism*, which licenses talk about moral truths and moral facts in a non-cognitivist framework, can also appeal to this justification to defend the theoretical virtues (Blackburn 1993, Gibbard 2003). The basic move that quasi-realism makes is to adopt *deflationist* analyses of notions like truth, fact, and correspondence, according to which saying that (some declarative sentence) *S* is true, expresses a fact, or corresponds to reality is simply saying *S*. This allows us to interpret moral claims involving these notions as equivalent to moral claims that do not involve them, which are (according to non-cognitivism) simple expressions of the speaker's non-cognitive attitudes.

So consider how the quasi-realist might deflate the apparently realist claims the justification makes so that they are acceptable to non-cognitivism. He might begin by paraphrasing things to eliminate the appeal to moral facts: instead of saying that the unitary hypothesis is more likely to entail *the fact that* causing injury is wrong, *the fact that* failing to reciprocate is wrong, and *the fact that* breaking promises is wrong, we can simply say that the probability that causing injury is wrong, failing to reciprocate is wrong, and breaking promises is wrong is higher given the unitary hypothesis. Framed this way, the justification only makes the following assumptions: (1) that we have credences (degrees of belief) in moral claims, including conditional credences in moral claims on other moral claims, and (2) that these credences are governed by the norms that govern probabilities. With these two assumptions, we can still show, on Bayesian grounds, that we ought to have a higher credence in some simpler, deeper, and potentially

much more revisionary theory T_1 over some other theory T_2. And given that quasi-realism is already able to capture so much of moral discourse that once seemed proprietary to realism, one might think that it can easily capture the two assumptions I listed above.

Although a detailed assessment of this strategy would go beyond the scope of this paper, I want briefly to call its viability into doubt. First, one might doubt that quasi-realism can even capture the idea that we have *degrees* of belief in moral claims. As Michael Smith (2002) points out, non-cognitive attitudes have only one dimension along which they can vary, strength: I can disapprove of one act more intensely than I disapprove of another. And this variation maps onto variation in *how seriously wrong* we judge acts to be, rather than *how confident* we are in our judgments that those acts are wrong: I believe that murder is more seriously wrong than theft in virtue of my disapproving of the former more intensely than the latter. So there simply does not appear to be any dimension of variation in our non-cognitive attitudes that could track our degree of belief in moral claims.

Of course, some philosophers have argued that Smith is mistaken. On the account proposed by Nicholas Makins (2022), for example, someone's degree of belief in a moral claim can be captured by the degree to which he is unambivalent in his non-cognitive attitudes. Even if we can give a non-cognitivist account of degrees of moral belief *simplicter*, however, the argument requires that there be such a thing as degrees of belief in some moral claim *conditional* on some other moral claim: my credence that injuring someone is wrong, not reciprocating is wrong, breaking promises is wrong, and so on *given* the unitary hypothesis must be higher than my credence *given* the pluralistic one. And it is unclear if the non-cognitivist account of degrees of moral belief can accommodate the notion of degrees of moral belief conditional on other moral claims.

Third, even if it could, it's unclear why whatever plays the role of degrees of belief in moral claims would be governed by the same norms that govern probabilities. The justification used Bayes's rule as a formula for updating our credences when we encounter new evidence: our new credence in some proposition should be the product of the old credence and the likelihood of the evidence on the proposition, divided by the total likelihood of the evidence. There are good reasons why our degrees of belief should be regulated by this formula; in the absence of further arguments, however, there is no good reason why (say) the degree to which our non-cognitive attitudes are unambivalent should be regulated by it.

Of course, none of this is meant *conclusively* to preclude the possibility that

quasi-realism could employ the justification for virtues like simplicity and depth that I claimed as proprietary for realism. Quasi-realists have shown great philosophical ingenuity in making available to non-cognitivism areas of moral discourse that were once considered proprietary to realism, and perhaps they could find a way to render the justification that I have claimed as proprietary to realism acceptable to non-cognitivism. But given the significant obstacles that they must surmount in order to do so, I am pessimistic that they can.

3 Anti-realism and the theoretical virtues?

So far, I have argued that moral realists have a proprietary (although defeasible) justification for prizing theoretical virtues like consistency, simplicity, and depth: such virtues are conducive to the truth of moral theories. The question now is whether anti-realists, in my sense, are justified in prizing them. Note that anti-realists cannot justify appealing to them on the same grounds on which realists appeal to them; after all, anti-realists deny that moral claims are true in a way that could do such philosophical work. As Jonathan Bennett (1998, 18) writes: "What is so bad about [inconsistency]? For the realist the answer is easy: If a morality is inconsistent then it is not true. The non-realist, who denies that any morality can be true, must answer differently." We might agree with him not just about consistency, but about the other theoretical virtues too.

In this section, I want to consider some justifications available to an anti-realist that might allow him to reinstate the theoretical virtues. These justifications fall into three types. First, one natural alternative is to try to secure their status as *pragmatic* virtues, features that we have practical reason for wanting our moral theory to have: it may simply be easier or more convenient, for example, if our moral theory has certain features. Second, there might be *moral reason* to value certain features. Finally, we might think that we can justify caring about these features based on *personal preference* alone. I'll call these *pragmatic*, *moral*, and *personal* justifications of the traditionally theoretical virtues.

In this section, I'll argue that these justifications fail. I'll do this by considering and arguing against particular pragmatic, moral, and personal arguments for caring about the theoretical virtues. Although this won't be anything like an exhaustive list of such arguments, their failure gives us reason for pessimism about justifying a heavy reliance on the theoretical virtues if anti-realism is correct.[4]

[4]What about moral error theory? After all, error theory is a form of anti-realism, and it leads to a normative view—moral nihilism—that is supremely consistent, simple, and deep. Although this is true, error theorists do not accept nihilism *because* it has these virtues. Rather, they do so

3.1 Pragmatic justifications

First, consider pragmatic justifications of the theoretical virtues. There seems to be an obvious practical reason to prefer consistent and simple moral theories: they're more convenient to use. In the case of consistency, the practical reason might seem especially strong: if I have a moral theory that says both that an act is right and that it is wrong, I cannot act on that theory, and I'll face practical paralysis. This is extremely inconvenient, so I have strong practical reason to reject inconsistent theories. In fact, as Allan Gibbard (1995) argues, one might think that a moral theory simply fails at its function if it is inconsistent, since it does not offer clear guidance on how to act.

There is a similar line in philosophy of science, where some argue that features like simplicity are pragmatic virtues of scientific theories: we are justified in preferring theories that are simpler not because simplicity is conducive to truth, but just because simpler theories are easier to use. As Bas van Fraassen (1980, 88) writes,

> the answer is that the other virtues claimed for a theory are *pragmatic* virtues. In so far as they go beyond consistency, empirical adequacy, and empirical strength, they do not concern the relation between the theory and the world, but rather the use and usefulness of the theory; they provide reasons to prefer the theory independently of questions of truth.

The case might seem even stronger when we consider that the starting point for moral theorizing is our ordinary, pre-theoretic moral judgments, since those judgments are quite messy. Moral philosophers have noted, for example, that commonsense morality includes a number of basic concepts, which cannot be unified at any deeper level: avoidance of harm, promotion of the general good, fairness, special obligations, and so on (Nagel 1979). And these basic moral concepts are often highly contested, so not neatly or uncontroversially associated with any non-moral property: it is notoriously difficult to say what fairness or harm, for example, consists in. This means that any set of moral principles that capture commonsense morality will likely be quite complex and fairly shallow.

Beyond this, commonsense morality often renders inconsistent verdicts in particular cases. For one, because it recognizes many different kinds of duties,

---

because they believe, for reasons having nothing to do with the theoretical virtues, that moral language makes systematically false presuppositions, such as the presupposition that there are categorical reasons for action (Joyce 2001, ch. 2). So error theorists are not engaging in radical moral theorizing, in my sense.

these duties occasionally conflict, generating a moral dilemma. And we cannot always resolve these inconsistencies simply by pointing out that these are *prima facie* duties, so that one of the conflicting duties might outweigh or override the other, as Ross thought; in some case, the conflicting duties seem genuinely counterpoised. Take, for example, the famous case of Sartre's student, who had a duty of patriotism to join the resistance, and a conflicting, equally strong duty of filiality to stay and home and take care of his aging mother. The student might judge both that he should join the resistance and that he should not join the resistance. For another, ordinary moral agents are susceptible to *framing effects*, forming different moral judgments about identical situations that are merely presented differently to them. As one example, people tend to view a policy that is described as having a 90% chance of success as safe, while viewing the same policy as risky when it is described as having a 10% chance of failure. Hence we might judge, on separate occasions, both that we should and that we should not enact the policy. Given how inconsistent, complex, and shallow commonsense morality is, we might think that a revisionary theory that has the traditional theoretical virtues would be more convenient to use for generating verdicts about the cases that we are likely to come across.

There are several objections to this attempt to reinstate theoretical virtues as pragmatic virtues of a moral theory. One objection, which I won't develop in detail, is simply that pragmatic reasons simply do not seem like the kinds of things that moral reasoning should respond to—at least, not in this way. After all, the ultimate goal of moral reasoning is to allow us to determine what we are morally required to do on particular occasions. And it seems counterintuitive that claims about what we are morally required to do should depend for their justification on the pragmatic consideration that the theory that implies the moral verdict is easier to use than commonsense morality: if I believe that it is permissible to kill one in order to save five, it does not seem to be even a partial justification of my belief to say that the theory that implies that verdict is easier to use than commonsense morality is.

A second objection, which I will also mention briefly, is that even if consistency, simplicity, and depth contribute to the ease of use of a moral theory, it is still unclear how we should trade them off against goodness of fit. Perhaps, when two theories are equally revisionary, the theory that is more theoretically virtuous will be preferable to the other theory. But in the case of two theories that are revisionary to different degrees, should we prefer the more revisionary, but simple or deep, theory to the more conservative, but complex or shallow, one? It is unclear how we would even go about answering the question. In the

case of moral realism, we saw that we can evaluate a tradeoff between goodness of fit and the theoretical virtues in terms of a single currency, that of how much having each feature makes a theory more likely to be true. In contrast, if moral realism is false, we have no clear way of assessing how much goodness of fit we should be willing to sacrifice for a gain in ease of use. For all that the argument shows, perhaps we are not allowed to sacrifice any degree of fit for a gain in the theoretical virtues. If that is so, then we could only use these virtues in the very limited role of a tiebreaker between two theories that conform equally well to our existing judgments.

A final objection, which I want to develop in greater detail, concerns the claim that the messiness of our ordinary moral judgments translates to practical inconvenience. Perhaps the assumption is that people use commonsense morality as a *decision procedure*: a set of rules that we consciously apply to derive verdicts about particular cases. If this assumption were true, then given the messiness of commonsense morality, there would be pragmatic reason to prefer a revisionary moral theory. But given that ordinary people typically form moral judgments in everyday contexts effortlessly, the claim that our ordinary moral judgments are practically inconvenient is manifestly false, and the assumption behind it reveals a mistake about how ordinary moral judgment works. Rather, work in moral psychology shows that our ordinary moral judgments are typically the deliverances of a fast, unconscious, and affective system for forming judgments, what psychologists call *System 1*, rather than of conscious reasoning, *System 2* (Haidt 2007). Despite the effortlessness of System 1, its workings should not be thought of as simple; one of the takeaways of recent psychology is that it is a highly complex system sensitive to a wide range of features, in ways that are often incredibly difficult to formalize (Railton 2014). This is why, despite our facility at making ordinary moral judgments, a systematization of those judgments might be highly complex.[5] Now, this is not to deny that ordinary people form moral judgments effortlessly in *every* case that they encounter; dilemmas and other forms of moral indeterminacy do exist in everyday life. My point is just that in most of the morally significant cases that ordinary people regularly

[5]As one example of this, consider our reactions to the "footbridge" version of the trolley problem, in which we are presented with the option of pushing one person to his death in order to save the lives of five others. The vast majority of people think that it would be wrong to push the person, but there's no definitive explanation for why we think this: whether we should explain that judgment in terms of a rule against using others as means (Thomson 1985), or initiating a harm (Foot 1978), or physical assault (Mikhail 2011), or causing a harm when that harm is more causally proximate than the good brought about (Kamm 1989), and so on.

face, they have no trouble forming a judgment quickly.

Given the ease with which we ordinarily form moral judgments, a pragmatic argument for the theoretical virtues has a high bar to meet. And when we turn to revisionary moral theories, it becomes apparent that their theoretical virtuosity does not allow them to clear this bar. Even if we have a consistent, simple, and deep moral theory, we might have to do extensive calculation to determine what verdicts it delivers about specific cases. Take the case of hedonic utilitarianism, which is a paradigmatically consistent, simple, and deep moral theory. In order to apply the principle of utility to a specific case, however, one would have to calculate the consequences of many different acts, which might be forbiddingly complex. In fact, utilitarians since Mill have explicitly advised us *not* to treat the theory as a decision procedure, instead offering a range of easier-to-employ secondary principles that lead us, on average, to maximize utility. In general, given these objections, there does not seem to be a strong pragmatic justification to weighing consistency, simplicity, or depth heavily.

3.2 Moral justifications

Beyond pragmatic reasons, one might think that there are moral reasons for wanting our set of moral commitments to exhibit certain features. Here, I want to consider an argument made by Shelly Kagan (1991) for rejecting "dangling distinctions" in our moral theory: distinctions that we consider to be morally relevant, but whose relevance is not explained by more general distinctions that we recognize as morally relevant. Such distinctions are arbitrary, hence ought to be rejected.

Kagan (1991, 13f) writes, for example,

> Perhaps a slaveholder might find that a principle which distinguished according to skin color yielded intuitively correct judgments about when a gentleman is morally required to aid someone being whipped, and when he is not. Merely having found the distinction underlying his intuitions is not sufficient to justify it.

So we recognize that any distinction based on skin color is a dangling distinction, since there is no more general distinction one could correctly appeal to that would explain the moral relevance of skin color. And because it is dangling, we must conclude that it is not really morally relevant.

Kagan deploys this as part of an argument for a maximizing form of consequentialism, which rejects the relevance of the distinction between doing and letting happen. (More precisely, he argues that we must accept *either* maximizing

consequentialism *or* a minimalist view on which morality makes no substantive demands on us.) Because, according to Kagan, we cannot justify the relevance of such a distinction by deriving it from any other that we consider to be morally relevant, we should treat the distinction as irrelevant, thinking that letting harm happen is just as morally wrong as doing the harm oneself.

Now, it's unclear what Kagan's own reasons for rejecting dangling distinctions are. Perhaps they have to do with truth-conduciveness: we might think the inclusion of a dangling distinction makes a theory less likely to correspond to how things are, for reasons similar to the ones we discussed in §2. But we might imagine rejecting them on other grounds. We might, for example, reject them on moral grounds, treating the ban on dangling distinctions as expressing an Aristotelian conception of fairness, the injunction to treat like cases as like. In Kagan's version of this injunction, we must assign two acts that differ only in a single factor the same moral status unless we can justify treating that factor as making a difference. Otherwise, the thought goes, we would be treating the agents whose actions we are assessing unfairly, which is morally wrong. For example, if I judge Tom harshly for breaking a promise to a friend, then I must judge Jane equally harshly for doing so, assuming other relevant factors are the same; if I do not, I would be treating them unfairly.

The ban on dangling distinctions requires a theory with explanatory depth: if there are no brute moral differences, then we cannot treat some distinction as morally relevant unless there is an explanation for why we can do so, and we cannot treat any distinctions used in the explanation as morally relevant unless there is a *further* explanation for why we can do so, and we cannot treat any distinctions used in *that* explanation as morally relevant, and so on. But rejecting dangling distinctions would also secure something like simplicity in our moral theory. After all, consider how justifying a moral difference works on Kagan's model: we justify the moral relevance of some property N_1 by showing that everything with that property also has a broader property N_2, whose moral relevance we justify by showing that everything with that property also has a broader property N_3, and so on. At the bottom, we have a small set of properties whose moral relevance is self-evident. Furthermore, if we cannot justify taking some property N to be morally relevant, we have to reject its moral relevance, thereby no longer accepting moral principles that feature it. This reduces the number of foundational moral principles, as well as basic concepts, that we accept.

The argument, however, is flawed for several reasons. First, it's unclear that all dangling distinctions strike us as morally arbitrary, even when revealed as dangling. Of course, nearly all of us would reject a distinction based on skin color,

but it is hasty to generalize from this case. After all, there are many dangling distinctions that most of us accept, even under reflection: the distinction between self and other, between doing and allowing, between friend and stranger, and so on. In the case of these dangling distinctions, our typical reaction is not that they need but lack justification, hence must be rejected; rather, it is that they do not need justification in the first place. Of course, some philosophers attack such distinctions because they are unsupported, but it seems unclear why dangling distinctions should be bad unless we *already* cared about something like simplicity or depth in our moral theory. For that reason, relying on a general ban against dangling distinctions in arguing for those features may simply beg the question.

Second, even if we do accept a general principle banning dangling distinctions, it's unclear whether our commitment to such a principle is stronger than our commitment to specific moral judgments that the principle conflicts with. Consider the distinction between passively allowing harm to befall an innocent person and actively harming an innocent person. Suppose that we discover that the distinction is a dangling one, as Kagan thinks: that the contrast between allowing and doing harm is not a special case of some more general contrast that has moral relevance. If we reject dangling distinctions, then we must conclude that the act of intentionally drowning an innocent person is no worse than passively allowing that person to drown. At the same time, most of us have a strong intuition that conflicts with this conclusion: someone who allows another person to drown is cowardly or callous, but someone who intentionally drowns another person is a monster. And it does not seem obvious that our commitment to the conclusion is stronger than the strength of the conflicting intuition, or that learning that the intuition relies on a dangling distinction would weaken it. If we are more committed to the intuition about the particular case than to the general principle, then we should simply reject a wholesale ban on dangling distinctions, circumscribing it to the point where it cannot support simplicity and explanatory depth as well.

And third, on pains of regress, the process of justification that Kagan assumes entails that there is at least one dangling distinction, a morally relevant property that is not subsumed by a more general morally relevant property; Kagan (1991, 14) himself notes that justifications "have to come to an end *somewhere*." As a consequentialist, for example, Kagan takes the distinction between maximizing the good and failing to maximize the good as brute. But given that this moral difference is brute, it is unclear why we cannot appeal to brute moral distinctions more generally. We need some reason to think that less is better than more in

the case of dangling distinctions, and any reason that one could offer for this seems suspiciously similar to an appeal to simplicity, which would also beg the question.

Now, the argument might support *some* degree of concern for simplicity and depth: there may be some distinctions that we are inclined to reject once their lack of justification is revealed, and rejecting those distinctions would push us in the direction of a more unified moral theory. But the question is how far in that direction it would push us. Given that radical moral theorizing requires a high degree of emphasis on the theoretical virtues, enough to deliver highly revisionary theories, it is unlikely that a partial rejection of dangling distinctions would justify radical moral theorizing.

3.3 Personal justifications

Finally, beyond any pragmatic or moral reasons for wanting consistency, simplicity, and depth in our moral theory, we might want these things simply as a matter of *personal preference*. We might want these things simply because we like symmetrical architectonics or desert landscapes. Or, if we think that our moral commitments are expressions of our psychology, we might value a sense of psychological integrity. As Jonathan Bennett (1998, 21) writes, defending his own taste for radical theorizing,

> From a non-realist standpoint, I can explain my pursuit of high generality. As a personal matter I want to be guided by rather general moral principles. This desire is neither extractable from the concept of morality nor based on insight into the structure of the real. It seems to come from my wish to be whole and interconnected in my person, so that I can understand some of my attitudes as parts or upshots of other more general ones.[6]

Of course, different people differ in whether they prefer things like consistency, simplicity, and depth over goodness of fit. Philosophers are somewhat notorious for having systematizing impulses, so the mere fact that Bennett prefers unified moral theories to particularistic ones is not strong evidence that most ordinary people would too. It is entirely possible that those who have a strong preference for the theoretical virtues are in the minority, and that most ordinary people would favor a moral scheme that conforms well to our existing moral judgments.

[6]Though cf. Nietzsche's remark on "the will to a system": "I mistrust all systematizers and avoid them. The will to a system is a lack of integrity." (*Twilight of the Idols*, I 26)

This argument sacrifices offense for defense: one is unlikely to persuade anyone who does not share the same preferences to accept a theory that satisfies them, but at the same time, one's own engagement in radical moral theorizing can no longer be criticized, since it is a matter of taste. We are perfectly permitted to find certain acts morally right or wrong on the basis of the preferences, perhaps idiosyncratic ones, that we have; but we must also recognize that we may not be able to persuade anyone else that those acts are right or wrong.

What is wrong with such a limited defense of radical moral theorizing? Well, morality is objective in scope: if an act is morally right, then everyone in the right circumstances has reason to perform that act. Furthermore, as many philosophers have stressed, morality comes along with the sanctions of the reactive attitudes: when someone does something wrong, reactions like indignation or blame are thereby warranted toward him on the part of others. And there seems to be something strange about the idea that we could justifiably form beliefs about what other people have reason to do, and about when we are permitted to feel blame or indignation toward them, on the basis of personal preferences that we cannot justify to them.

There are several ways to flesh out this intuition. One way would be in terms of internalism about reasons (Williams 1979), according to which someone has reason to do something only if deliberation on her subjective motivations could lead her to accept such a claim. A second way would be in terms of a public-justification constraint on moral norms (Gaus 2011), according to which a moral claim is authoritative to someone only if she has sufficient reason (understood on an internalist model) to accept that claim herself. My preferred way, however, is in terms of a principle that David Enoch (2013, ch. 2) mentions: that in cases of conflict between our preferences and others'—at least when the decision affects others—we should step back and adopt an impartial solution to the conflict. If everyone else in the group wants to go to the park together while I want to watch a movie together, it would be wrongfully self-assertive for me to insist on my preference. Rather, I should abandon that preference given that my preference concerns what everyone will do, and that everyone else has a conflicting preference. Similarly, if a personal preference for simplicity leads me to accept a moral theory that makes revisionary verdicts—say, utilitarianism—it would be wrongfully self-assertive for me to insist on those verdicts, condemning others for not killing one to save five. Rather, I should abandon that preference and the verdicts that it supports given that those verdicts concern what other people should do, and that the vast majority of those people do not share that preference.

More generally, there is something intuitively strange about allowing our judgments about something as serious as morality to be determined by preferences that do not require justification. To the extent that we think our moral judgments should be based on something firmer than that, a justification for favoring the theoretical virtues in terms of personal preference fails.

Conclusion

I've argued that radical moral theorizing, which assigns a heavy role to theoretical virtues like consistency, simplicity, and explanatory depth in moral inquiry, is most naturally compatible with moral realism, broadly construed. After all, one natural justification for valuing such features is that they are conducive to truth, understood as correspondence to the moral facts. Anti-realists cannot appeal to such a justification, so they must justify caring about these features for other reasons: pragmatic, moral, or personal. Although anything like an exhaustive discussion of these reasons for engaging in radical moral theorizing is impossible, I've argued that some natural defenses of radical moral theorizing along these lines fail. We have reason for pessimism about the prospects for a highly theory-driven approach to moral inquiry if moral realism is false.

This is not to say that, if realism is false, there will no longer be a role for moral theorizing; even if we can no longer do *radical* moral theorizing, there will still be room for more *conservative* moral theorizing. After all, consider that some of the ambitions of moral theorizing are not revisionary: we theorize partly to explain moral phenomena and to understand the connections between different moral concepts. And even if the theoretical virtues do not matter as much in theorizing, there will still be some pressure toward revision. There might be local areas where the inconsistencies in our moral judgments become unwieldy, or where we recognize that our judgments depend on distinctions that we do not endorse under reflection; in these cases, we might have to reject some of our pre-theoretic judgments. What I deny is simply that this pressure will lead us to accept moral theories that reject large swaths of those judgments in the name of the theoretical virtues. We might want a straighter intersection here or there, but nothing like a Haussmannian demolition and reconstruction of entire neighborhoods to produce conformity to an overall grid. If moral realism is false, then moral philosophy may have to leave everything roughly as it is.

References

Jonathan Bennett. *The Act Itself*. Clarendon Press, 1998.

Simon Blackburn. *Essays in Quasi-Realism*. Oxford University Press, 1993.

Julia Driver. "Moral Theory." In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. 2022.

David Enoch. *Taking Morality Seriously: A Defense of Robust Realism*. Oxford University Press, 2013.

Philippa Foot. "The Problem of Abortion and the Doctrine of Double Effect." In *Virtues and Vices*, pages 253–255. Blackwell, 1978.

Steven French. *The Structure of the World: Metaphysics and Representation*. Oxford University Press, 2014.

Gerald Gaus. *The Order of Public Reason*. Cambridge University Press, 2011.

Allan Gibbard. "Why Theorize How to Live with Each Other?" *Philosophy and Phenomenological Research*, 55(2):323–342, 1995.

Allan Gibbard. *Thinking How to Live*. Harvard University Press, 2003.

Jonathan Haidt. "The New Synthesis in Moral Psychology." *Science*, 316:998–1002, 2007.

Brad Hooker. *Ideal Code, Real World*. Oxford University Press, 2003.

Brad Hooker. "Theory and Anti-Theory in Ethics." In *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams*. Oxford University Press, 2012.

Michael Huemer. "Revisionary Intuitionism." *Social Philosophy and Policy*, 25: 368–392, 2007.

Richard Joyce. *The Myth of Morality*. Cambridge University Press, 2001.

Shelly Kagan. *The Limits of Morality*. Oxford University Press, 1991.

Frances Kamm. "Harming Some to Save Others." *Philosophical Studies*, 57:227–260, 1989.

Nicholas Makins. "Attitudinal Ambivalence: Moral Uncertainty for Non-Cognitivists." *Australasian Journal of Philosophy*, 100:580–594, 2022.

John Mikhail. *Elements of Moral Cognition*. Cambridge University Press, 2011.

Thomas Nagel. "The Fragmentation of Value." In *Mortal Questions*. Cambridge University Press, 1979.

Derek Parfit. *Reasons and Persons*. Oxford University Press, 1984.

Peter Railton. "The Affective Dog and Its Rational Tale: Intuition and Attunement." *Ethics*, 124:813–859, 2014.

W. D. Ross. *The Right and the Good*. Clarendon Press, 1930.

Geoffrey Sayre-McCord. "The many moral realisms." *Southern Journal of Philosophy*, pages 1–22, 1986.

T. M. Scanlon. *What We Owe to Each Other*. Harvard University Press, 1998.

Michael Smith. "Evaluation, Uncertainty, and Motivation." *Ethical Theory and Moral Practice*, 5(3):305–320, 2002.

Elliott Sober. *Ockham's Razors: A User's Manual*. Cambridge University Press, 2015.

Judith Jarvis Thomson. "The Trolley Problem." *The Yale Law Journal*, 94:1395–1415, 1985.

Mark Timmons. *Moral Theory: An Introduction*. Rowman and Littlefield, 2013.

Bas van Fraassen. *The Scientific Image*. Oxford University Press, 1980.

Mark van Roojen. *Metaethics: A Contemporary Introduction*. Routledge, 2015.

Bernard Williams. "Internal and External Reasons." In Ross Harrison, editor, *Rational Action*. Cambridge University Press, 1979.