Moral Overfitting

Audrey Powers

Forthcoming in *Philosophical Studies*
(Penultimate draft – please cite published version)

Abstract

This is a paper about model-building and overfitting in normative ethics. Overfitting is recognized as a methodological error in modeling in the philosophy of science and scientific practice, but this concern has not been brought to bear on the practice of normative ethics. I first argue that moral inquiry shares similarities with scientific inquiry in that both may productively rely on model-building, and, as such, overfitting worries should apply to both fields. I then offer a diagnosis of the problems of overfitting in moral inquiry and explain how our current practice seems worryingly susceptible to such problems. I conclude by giving suggestions for how we might avoid overfitting when doing normative ethics.

1 Introduction

This is a paper about model-building and overfitting in normative ethical inquiry.

Ethicists are generally not too worried about these topics. The ultimate project of first-order normative ethics as traditionally conceived is to discover the moral laws of the universe. Insofar as we are engaged in this project, we want to find exceptionless, universal principles that govern or give precise rules for the ways the moral facts are distributed in the world.

These commitments are often implicit in but essential to our practice (Rosen, 2017b). When engaging in ethical inquiry, we want to find the rule in virtue of which, e.g, it's wrong for me to tell a particular lie. And, we think, this had better be the same rule in virtue of which it's wrong for you to tell a particular lie, and in virtue of which it's wrong to lie in general. If that rule admits of exceptions or counterexamples, we haven't done our job well – we need to revise our rule or come up with a different one. That is to say: we want to discover the moral laws. And so long as we are not moral particularists (e.g., Dancy, 2004), which most of us aren't, we think that there really are such true, universal generalizations about the moral to be discovered – that is, that moral laws exist in some meaningful sense.

Such moral laws, however, are hard to come by. I do not think that most ethicists who are engaged in the law-seeking project of moral inquiry would bet that we have successfully discovered any laws as of yet. And the ethicists who would not take that bet might also be engaged in an alternative project of building moral models, analogous to scientific models. These ethicists might give up some of the aims or methods of the traditional project. But they might, via modeling, gain knowledge of the moral laws, and gain access to virtues like accuracy, predictiveness, and action-guidingness. Or so I will argue.

In this paper, I talk about what such a style of moral theorizing looks like. I also talk about the perils of this approach. Namely, if we are building moral models, we need to be wary of overfitting these models to our data. Moral models should not be overly flexible, and should not adhere to the data too closely or in an unprincipled manner. Such worries are well-established in philosophy of science (Forster and Sober, 1994, Hitchcock and Sober, 2004), scientific practice (Hawkins, 2004, Ying, 2019), and philosophical methodology in general (Hanson, 2002, Alexander and Weinberg, 2014, Weinberg, 2017, Williamson, 2021, 2024), but have not yet been articulated in regard to normative ethics. I argue that moral modeling shares similarities with scientific modeling such that overfitting worries should apply to both fields of inquiry, and I provide some preliminary solutions for problems of overfitting in ethics.

In this way, my project is distinct from other methodology-related questions in ethics such as concerns about the method of reflective equilibrium (e.g., Singer, 2005, Kelly and McGrath, 2010, McPherson, 2015) or about the use or relevance of cases, thought experiments, and intuitions (e.g., Unger, 1996, Cappelen, 2012). I am not so concerned with the accuracy or inaccuracy of our moral judgments, or how we might make these things better. I am instead concerned that there are methodological lessons that we might learn from the sciences that, thus far, we have been neglecting, and that our practice of normative ethics may be made worse off as a result of this neglect.

To this end, here is how things will go in this paper. I first (§2) make a case for the practice of normative ethics as involving modeling, via an analogy with modeling in the natural sciences. I then (§3) offer a diagnosis of the problems of overfitting in this model-based moral inquiry, and explain how our current practice of normative ethics seems worryingly susceptible to such problems. Finally (§4), I give some preliminary suggestions for how we might avoid overfitting when engaging in moral theorizing.

2 Normative Ethics and Modeling

Normative ethical inquiry is sufficiently like scientific inquiry such that the scientist's tools are good for the ethicist. More specifically, a scientific tool that is good for the ethicist is modeling – ethicists may build moral models that are usefully accurate, predictive, and action-guiding. I make these claims to set up the overfitting worries I bring up in the following section. It wouldn't make sense to be worried about overfitting if our practice didn't permit of it, so this is an argument that our practice permits of it.

2.1 Normative Ethics and Scientific Practice

We will get off on the right track by considering some similarities between philosophical practice and scientific practice. Some think of philosophy in general as continuous with or as a branch of the natural sciences, and therefore permitting of progress by use of scientific tools. This is not an uncontroversial claim, but it is not unheard of, either. We may associate such a view with Quine (1957) or Russell (1912), and more recently, e.g., Paul (2012) and Emery (2023) argue that scientific tools and methods are appropriate and necessary for use in philosophical inquiry.

Of course, these philosophers are talking in large part about metaphysics as a good locale for the use of scientific tools. What about ethics? Is scientific practice applicable to ethical inquiry? There are reasons to think so, especially if we fill out the subject matter of 'ethical inquiry' in certain ways.[1] The view of ethical inquiry as like or continuous with science is, again, not uncontroversial, but not unheard of – Dworkin calls such a view the "natural model," on which we discover the truths of morality like we discover the truths of science (1978, p.160).

Some metaethical commitments lead us towards such a view. If we're moral naturalists, we think the moral facts in some meaningful sense *are* natural facts. It would be strange, then, if our strategies for discovering the natural facts had no bearing on discovering the moral facts. In fact, a common naturalist characterization of what moral facts are is that they are the kind of things science tells us about (see, e.g., Darwall et al., 1992, Smith, 1994, Shafer-Landau, 2003, Copp, 2007 for discussion). If we're non-naturalists, we think moral

[1] About the subject matter of ethical inquiry: I am assuming moral realism for simplicity's sake, and because it seems to me that most ethicists assume it as well. I don't think this assumption is exactly necessary for my project – I think there are interesting things to be said about what we're doing if we end up building moral models where no actual moral facts exist – but I don't have space to say them here.

facts aren't natural facts, but still generally think that there are tight metaphysical relations between the moral and the natural – perhaps natural facts fully (Leary, 2017, Berker, 2018) or partly (Rosen, 2017a, Enoch, 2019, Fogal and Risberg, 2020) ground moral ones; perhaps there cannot be a change in the moral without a change in the natural (McPherson, 2019, Streumer, 2024). It would be strange, then, to think that our strategies for inquiry into the natural have no use in moral inquiry – or, at least, the burden of proof is on those who would claim this.

But all of this only gets us so far. Maybe it's true that scientific tools are useful for the ethicist. But still we may ask: what are 'scientific tools' – do I need to claim here that there's one standard or correct scientific methodology, and say what it is? This is unnecessary for my purposes. What I'm talking about are individual tools that natural scientists actually use to accurately predict phenomena they care about, as I am hoping that such tools will help ethicists do the same.

The particular tool that I want to claim as a success, particularly in the special sciences – and that I argue we may adopt to achieve certain ends in normative ethics – is modeling. Generally, we agree that scientific inquiry has been wildly successful as a method of making useful and accurate predictions about the world (see e.g., Putnam, 1979). I take models to be non-law predictive tools (I'll say much more about what models are in §2.2), and I take them to be a means by which we have achieved this success.

To be clear, modeling is not the only success of scientific inquiry. Another success is the discovery of a small number of natural laws – these govern (on anti-humean views, e.g., Armstrong, 1983, Wilsch, 2020, Emery, 2022) or summarize (on humean views, e.g., Lewis, 1973, Loewer, 2012, Hicks, 2017, Loew and Jaag, 2018) the distribution of phenomena in the natural world. But discovering (exceptionless, universal) laws is difficult, and does not seem obviously possible or perhaps even desirable in some of the natural sciences. It is at the very least controversial whether there are laws of biology in the way that there are laws of physics (Hamilton, 2007) – it's just not obvious that there exist any true, exceptionless generalizations of the humean or non-humean sort that quantify over the subject matter of biology.

And yet we know a great deal about biological processes. We know, for example, a great deal about the structure and workings of cell membranes, even though we do not have a natural law for such structures or workings. What we have instead is the fluid mosaic model (Singer and Nicolson, 1972), which informatively represents such structures and workings

and allows us to accurately predict, say, when things outside the cellular environment may enter or be prevented from entering a cell. Regardless of the status of biological laws – if there are such laws and we just haven't discovered them yet, or if there are none – biologists have found very effective ways to continue doing biology.

But we also engage in modeling in sciences where we know that there exist laws. If there are any natural laws at all, surely some of these laws operate in the domain of fundamental physics (think, e.g., of Newton's laws of motion). But not all of what physicists do involves working with such laws. An enormous success of particle physics is the Standard Model for the classes of elementary particles – this model predicted the existence of the Higgs boson, and does exceptionally well when it comes to describing and explaining fundamental forces other than gravity. The Standard Model is clearly not itself a law. It is neither exceptionless nor universal, leaving some physical phenomena unaccounted for. But it gives us more information about particle physics than we would have if we just worked with the laws.

And if this is a win in physics, we could sure use some wins like it in normative ethics. As ethicists, we should very much like to have a model that gives us accurate information about particular moral facts as the Standard Model gives us information about particular microphysical phenomena. We want to know when it is wrong to tell lies, and to break our promises, and to turn the trolley, and to prioritize the well-being of a few of our loved ones over the well-being of many strangers, and so on, and so on. In possession of such moral facts, we may know how to act, and what are right and wrong things to do. Insofar as ethics is meant to be action-guiding, findings like these will help us. We also want explanations for why these moral facts obtain – and models may help us here too. In the following subsection, I say more about how models can do these things for us.

2.2 Moral Models

So, as scientists build models of the workings of cell membranes and fundamental particles, normative ethicists may build moral models. Here, I say more about the form and function of moral models.

Moral models, as I take them to operate, take as inputs descriptive facts about cases or scenarios and output predictions about moral facts.

This is different from a moral law: on standard conceptions of the metaphysics of moral

laws (e.g., Rosen, 2017a, Enoch, 2019, Fogal and Risberg, 2020), these laws take as inputs descriptive facts and output *actual* moral facts – they do some work to set what the true moral facts really are. They also help us to explain why the moral facts are what they are – for instance, on this picture, a particular instance of lying is wrong because of certain descriptive facts that make that instance a lie together with the moral law that lying is wrong.[2]

Moral models aren't like this. They might give us epistemic access to the actual moral facts, if they're any good, but they do no work to set these facts. Even a perfect moral model doesn't help to make it the case that a particular instance of lying is wrong. But moral models can tell us about the world with a high degree of accuracy. If we're successful at this as scientists are in their modeling, we will be very successful. Presumably, we wouldn't want to discover the actual moral facts or laws merely to have discovered them – we would want to use these discoveries to see whether specific actions would be right or wrong. We care a great deal about what the moral facts are. And with a good model, we may access these facts. By means of modeling, we may learn about how to act morally, how to live our lives, and so on, without having discovered the moral laws.

And because moral models don't set moral facts, they also don't explain moral facts in the same way a moral law does. However, models can help us with explanation, if they hew closely enough to the actual moral workings of the world. If a moral model is good – if its predictions align with the actual moral facts[3] – we know that the moral laws cannot be such that they make different moral facts obtain than what the model predicts. In fact, we know the moral laws are such that they make these actual moral facts obtain. So we can posit moral laws accordingly. Modeling can thereby bring us closer to knowledge of the moral laws.

This is not to say that moral models must be true in the sense that moral laws are true. We may think of models as things we use

> to gain understanding of a complex real-world system via an understanding of simpler, hypothetical system that resembles it in relevant respects. (Godfrey-Smith, 2006, p.726)

[2]Not all accounts of the metaphysics of moral laws have laws working to set moral facts in this manner – according to Berker (2018), moral laws are perfect, exceptionless summaries of explanatory relations between descriptive and moral facts. Moral models are unlike such laws as well. These laws, metaphysically speaking, are descriptive rather than predictive. And these laws are by definition true, being complete summaries of the moral world.

[3]It's not clear how (or perhaps even whether) we can know a model's predictions align with moral facts – this depends on what we want to say about our epistemic access to the moral facts. In this way, there's a potential disanalogy between moral and scientific modeling. It's intuitively clearer how we can check scientific models for predictive accuracy. But I take the predictiveness of models to be a separate issue from our ability to check their predictions against actual facts, and set aside epistemic issues of how we might check models' predictions for the purposes of this paper.

Good models might (and often will) be in some sense false – they oversimplify, or do not fully explain, or make use of entities that do not exist. But false models may give us useful information about how to move forward in the modeling process, or, better, useful information about the world itself (Wimsatt, 1987). And this is desirable – when I'm trying to represent an incredibly complex system as a cognitively limited agent, idealizations will be useful for me (Potochnik, 2020). Since models may be false or idealized, we may even have a set of valuable models in a particular domain of inquiry that are mutually inconsistent.[4]

So – moral models take as inputs descriptive facts and output predictions about moral facts, and are not moral laws. This is still pretty abstract. Here's a slightly less abstract question: how do these concerns relate to our practice of normative ethics? And some even less abstract ones: what's an example of a moral model? What are examples of descriptive facts? I'll take these questions in reverse order.

Ethicists often work by first considering some descriptive facts, then feeding such facts into a model and examining the predictions that this model outputs (and testing those predictions against other data – I say much more on this last bit in §3). These descriptive facts might be imagined or actual. Here is a well-known and well-loved example of some descriptive facts:

> Trolley: A bystander sees an out-of-control trolley barreling towards five people working on the trolley tracks. The trolley will kill all five if it continues on its path. There is also a spur of track off to one side, on which one person is trapped. It is possible for the bystander to throw a switch and divert the trolley onto the side track, killing the one instead of the five. (Thomson, 1985)

Notice what isn't part of the set of descriptive facts here: any actual facts about what the bystander may or may not or should or should not do. Such facts are moral rather than descriptive.[5]

Now, if such facts are what we're going to input into a model, here's an example of the kind of model we're going to feed them into.

Imagine I'm an actual-consesquence act utilitarian. I think that the moral law is that the right action maximizes actual utility (Singer, 1977). I disagree with foreseeable-consequence

[4] I remain relatively neutral on what models are – I take what I say here and throughout this paper to be compatible with the work on modeling cited above, as well as standard views in the literature on modeling, e.g., Cartwright, 1983, Morgan and Morrison, 1999, Weisberg, 2013, 2016.

[5] I borrow a standard conception of descriptive facts from Dunaway, 2014, Streumer, 2024: these are the facts that can non-accidentally be ascribed with a descriptive predicate (as opposed to a moral predicate for moral facts). This is compatible with whatever view we like of what *makes* facts descriptive rather than moral (see Streumer, 2017).

act utilitarians, who think that the moral law is that the right action maximizes utility as far as the actor can reasonably predict (Gruzalski, 1981). But, as an actual-consequence act utilitarian, I take the foreseeable-consequence act utilitarian's moral law as a moral model – when we're acting, if we're trying to maximize actual utility, of course we will choose the option that we foresee as maximizing utility.[6] I think that this model will give good predictions about the moral facts, given these commitments. It takes as input descriptive facts (the facts in this case being TROLLEY, as described) and outputs predictions about the actual moral facts (that it's right to turn the trolley). It might give us accurate predictions, if the moral fact of the matter really is that the bystander ought to turn the trolley. It might even help explain the moral facts, if it helps to make it clear that the rightness of turning the trolley has something to do with maximizing utility. But there also exists an actual moral fact of the matter that depends on what will maximize utility, and that obtains regardless of what anyone *thinks* will maximize utility, and regardless of what my model outputs.[7] So my model is not a law – it does nothing to bring about the moral facts, nor is it itself a true summary of all moral facts – but it is certainly useful to me when it comes to figuring out what I should do, and why I should do it.

As such: there's a good amount of modeling going on in normative ethics. Ethicists regularly work with pieces of theoretical machinery that i) output predictions about what we should do, or what is right and what is wrong, when given descriptive facts as inputs, and ii) are not themselves moral laws – even if such ethicists think that there are moral laws out there. That's the general picture of normative ethics as model-building. In the next section, I see what conclusions about methodology in ethics apply in light of this picture.

3 Moral Overfitting

One of the uses of moral models is to make accurate predictions about moral facts. One thing we know stands in the way of predictiveness in scientific modeling is overfitting. So: moral models, being relevantly similar to scientific models, shouldn't be overfitted either. In

[6] Actual-consequence act utilitarians generally explicitly take the foreseeable-consequence act utilitarian's purported moral law as a decision procedure. I want to be clear that what I mean by 'model' is not merely a decision procedure. Models may be used as decision procedures or components thereof, if people actually decide on such a basis. But not all models are or should be used in this manner, and we can imagine decision procedures, even those that deal with the moral, as not having the traits – i.e., idealization, simplification, taking as inputs descriptive facts and giving predictions about moral facts as outputs – that I point to as characteristic of moral models. ('Heads I'll flip the switch, tails I'll do nothing' is a decision procedure and not a model.)

[7] So long as morality is objective and independent of us – which, again, I'm assuming.

this section I describe overfitting as a particular concern for normative ethicists, show how it might stand in the way of accuracy and predictiveness, and argue that the historical and present-day practices of normative ethics do not do enough to guard against overfitting.

I take overfitting to be a problem in model construction. Williamson defines overfitting in terms of a model's "having too many degrees of freedom" such that "one can fit just about any data, but in a cheap way which typically brings no insight" (2021, p.79), and in terms of a model-builder's "willingness to add extra parameters to an equation until its curve goes almost exactly through all the data points" (2020, p.264). Weinberg writes of "having one's model become itself captured by spurious twists and turns in the data" (2017, p.258).

Though Williamson and Weinberg both write of overfitted models, they bring these concerns to the project of theory construction. When speaking of the data to which we might fit a theory, both refer to a category that includes, presumably, both the regular descriptive facts of the kind I describe earlier, but also our intuitive verdicts on particular thought experiments (Williamson calls this latter kind of data "*prima facie* evidence" (2024)). Weinberg wonders if defenders of justified true belief-style theories of knowledge ought not to worry so much about Gettier cases, as perhaps such implausible or uncommon or just plain weird cases and our intuitions about them are not the kind of data to fit a theory to. So our practices may make theories of knowledge worse. Williamson worries that, for example, innovations in dynamic semantics might turn out to be instances of overfitting theories to data by adding unnecessary parameters, if it turns out the data that drives these innovations are otherwise explainable (2024). So our practices may make semantic theories worse.[8]

These concerns apply to moral models in a distinctive manner. We care very much about accurate predictiveness in moral models: I want my model to correctly tell me whether I should keep a promise, regardless of whether it tells me anything about the nature of morality in general. We care differently or less about accurate predictiveness when it comes to epistemological theories: I'm not so invested in whether Smith really knows that the man who will get the job has ten coins in his pocket. I care about this merely as it helps me analyze knowledge. So insofar as overfitting makes models less accurately predictive, I should be particularly interested in avoiding overfitting in moral modeling.

In the moral case, we run the risk of overfitting models to data as we do elsewhere. When

[8]To be clear, I don't necessarily endorse Williamson and Weinberg's claims that these in particular really are instances of overfitting – I use them as potential examples.

I speak of 'data' in this context, I'm talking about input-output pairs: we input descriptive facts in hope that our models will output predictions about the moral facts. But we often have input-output pairs in mind prior to our construction of models. When I consider the descriptive facts of a situation in which a doctor who has the option to kill a healthy patient to provide organ transplants to five sick ones, I hope that any model I build will output the prediction that it would be wrong for the doctor to do so. So I fit my model to this input-output pair: I want to make it so that when I input these descriptive facts, the model indeed outputs the prediction that the doctor may not kill the patient.

Of course it is possible to fit a model to input-output pairs correctly. But it is possible to overfit a model to such pairs too. Let me say more about this latter possibility. What does an overfitted moral model look like? And in what manner is such a thing bad? To answer these questions, I'll look at what I take to be an example of an overfitted model in historical ethical inquiry – that is to say, I'll be picking on Kant.

To present-day readers of Kant, it seems not only that some components of his moral philosophy were objectionably sexist, but that these components were objectionably sexist by Kant's own lights.[9] There is something not only wrong but bizarre about Kant committing himself to the "superiority of the husband to the wife" while wondering if the inequality of the marriage contract "is in conflict with the equality of the partners" (MM 6:279), to say nothing of the equality presupposed by the humanity formulation of the categorical imperative: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (G 4:429).

I offer a diagnosis of this strangeness. This is an example of overfitting – adding ad-hoc parameters to one's model as a result of over-adherence to data. I do not mean to claim that what I give here is a description of Kant's own thinking when engaging in ethical inquiry, but I do give a description of the way such a moral model is overfitted.

The humanity formulation of the categorical imperative does not distinguish by gender, referencing "persons" in general rather than persons of any particular gender. As contemporary Kantians note, it just does not return the result that Kant thought it returned. If anything, it clearly tells us that the inequality of the marriage contract is morally wrong. But note some "*prima facie* evidence" available to Kant: individuals in his vicinity took the inequality of the marriage contract to be unremarkable, even morally correct. Say we took

[9]See, e.g., Langton, 1992, Kleingeld, 2019, among many others.

this kind of judgment to inform the desired output of a moral model. Then there would be an input-output pair – an input of descriptive facts regarding the social position of women in marriages, and an output prediction that this state of affairs is morally permissible – to which we might want to fit our model.

To do so, we might attach a new parameter to the categorical imperative. This new parameter might be the claim of the superiority of the husband to the wife. By adding this parameter, we could input descriptive facts into our model about the ways that women are subjugated within marriages, and the model would output the prediction that this state of affairs is fine, morally speaking.

Now, clearly we ought not to add this new parameter. And of course, one way – I think the best way! – for *us* as contemporary readers of Kant to diagnose the problem is to say, look, the output here is just plain wrong, as gendered inequality is morally impermissible. But Kant seems to have lacked epistemic access to this particular moral fact. So perhaps *he* could have noticed that there seem to be some methodological pitfalls that come with the addition of extra parameters to models. We should be wary of fitting models too closely to all data in general: accommodating certain inputs-output pairs (such as the transplant case, and the prediction of its moral wrongness) doesn't seem intuitively suspect, but accommodating certain other input-output pairs does, whatever it is that makes those other pairs bad. And we can note that some additions make models complex where they were previously simple, specific where they were general, and flexible where they were more rigid. We have good reasons to at least entertain the possibility that our machinery might be better without such additions.

This is to say, overfitting is not the only problem with a model like Kant's. But it is a sign of its problems – one way we could discover that such a model is bad is by noticing that it is overfitted. From a model-construction standpoint, we may identify a number of issues with such models.

First, overfitting encourages unprincipled choices in model-building. Adding new parameters to a model to accommodate *all* data has us run the risk of positing ad-hoc parameters, if we are not careful – as this Kantian one seems to be ad-hoc. As Hitchcock and Sober note, "the more one's background theory makes it easy to accommodate new data, the less the success of that theory in accommodating the data redounds to the credit of the theory" (2004, p.7, see also Popper, 1962) – that is, if we take this profligacy as giving us a good result, we

give a bad model too much credit.

Second, one can model just anything with an overfitted model. Williamson (2021) calls such models "cheap" and "uninformative." For instance, fitting a curve to randomly selected points does not tell one anything about the points, and makes one think there is a connection between such points to be usefully represented by a model when that is not necessarily the case. This Kantian model will have us think that our input is meaningfully connected to our output in a way that is represented by the model – that the fact that women are subjugated in marriages is morally innocuous in the manner that the model tells us. But of course this is false. Overfitted models will have us seeing such connections when they do not exist.

Third, overfitted models are not predictive, or not accurately predictive. If one is willing to revise one's parameters to accommodate all data, one runs the risk of trading the predictive power that comes with a more rigid model for the accuracy of, say, a curve that passes through all data points. Certainly not all accommodation of this sort is bad – good models should have something to say about existing data, and shouldn't be constructed *a priori* (see Howson, 1990, Hitchcock and Sober, 2004). But overaccommodation is bad – it makes it less likely that the curve will pass through future data points without further tweaks to the parameters. And this should have us worry about such a model's predictive power.

If such a model simply failed to predict, this would be bad – Popper (1959) critiques non-predictive theories as unfalsifiable and therefore unscientific. But it also seems possible that an overfitted model will predict, just wrongly (Hitchcock and Sober, 2004). Because of its extra parameter, our overfitted Kantian model predicts not only that inequality within the marriage is morally acceptable, but that gender-based reproductive inequality, income inequality, inequality in public life, and so on are acceptable. This is not what we want from moral models – if they are to be useful, they should be accurately predictive. Making a model more flexible to capture particular data might negatively impact the model's future predictive and explanatory power.

Fourth and last, overfitted models do not help us distinguish good data from bad. Datasets have problems – this is a sad fact about our world. Moral datasets certainly do – sometimes our *prima facie* evidence is just false, as we are not perfect predictors of moral facts. Consequently, we will sometimes be wrong about the truth value of the outputs to which we want to fit our models. Model-builders should help themselves to tools to find out these problems (see again Williamson, 2021). Some such tools are sufficiently rigid model components that

help to distinguish data that is worth modeling from data that is not worth modeling. If one's model admits of inputs in a certain format (i.e., as a polynomial equation admits of ordered pairs), when confronted with a relevantly different type of input, the model ought not necessarily just accommodate it (i.e., a polynomial equation does not admit of an ordered triple, and a principled equation-builder should not add an ad-hoc parameter to a polynomial that consists of a component to deal with ordered triples when they occur).

In the case of a moral model: perhaps one's model should admit of only certain kinds of inputs, depending on the moral facts one wishes to predict. Without model components that help us in this manner, we run the risk of treating good data and bad data the same way. Alexander and Weinberg (2014) call methodologies that are insufficiently rigid and therefore unable to help with the discovery of bad data "error-fragile" – overfitted models are unable to deal with bad inputs, and modelers are in danger of unnecessarily altering good models in order to account for data which (unbeknownst to the model-builder) are faulty. To return to our Kantian model: if it were more rigid and did not admit of alterations to accommodate data about existing gender inequality or attitudes about this inequality, and did not allow the relevant inputs, the outputs such a model would give would be different. We would not get the same prediction that widespread gender inequality is morally acceptable.

I do like to think we are doing better than Kant when it comes to avoiding overfitting our models to faulty input-output pairs in regard to gendered inequality – in part because we seem to be better than Kant at identifying outputs that are faulty in this manner. But I am not sure we are doing better than this in every aspect of our practice.

Our datasets must still have problems. It is natural to wonder what those reading contemporary analytic philosophy in hundreds of years would pick out as our errors as we pick out Kant's. It would be hubristic to suppose that we aren't blind to any unjustified social inequality, or to assume that our moral models deal with such inequality perfectly instead of containing ad-hoc components that allow us to attempt to justify our practices. Surely, like Kant, we presently fit our models to some input-output pairs for which we are simply wrong about the output.[10] The point is this: we are not in a moral or epistemic position to be

[10]It's hard to make good predictions about what current practices of ours might be regarded as horrific in the future. We can at least imagine, for example, our treatment of animals as being a candidate for such a practice, and we can imagine what our descendants might say about it (see, of course, e.g., Singer, 1975). But I don't need to point to specific cases here. I just want to note that people used to think [insert your least-favorite widespread historical atrocity here] was fine, and presumably we presently think the same about what might later be identified as historical atrocities.

confident that we can sort all good data from bad. And insofar as our dataset has problems, if we do not guard against overfitting models to problematic data, those problems will show up in our models as well.

This is particularly worrying because a great deal of work in ethics that is, first, paradigmatic of our professional practices, and, second, widely thought to be very good, or at the very least quite famous, deals with data in a way that runs overfitting risks. In normative ethics, it is standard to build a moral model, see what predictions that model outputs, and revise the model when its outputs are not as we would like them to be – usually, when they conflict with our *prima facie* evidence. But when we revise that model, we do not generally consider overfitting worries such as the ones that I present here, as such worries are not a significant feature of our methodology.

We can see this process particularly clearly in the literature on the trolley problem. A model is unable to output the moral predictions we might expect or want in regard to the descriptive facts about a particular trolley case that we input – therefore, we think, something is wrong with the model, and we must revise it so it will output different predictions (e.g., Thomson, 1976, Kamm, 2015). See Rosen on this standard practice:

> When we set out to explain why it's wrong to push the fat man, we seek a feature that distinguishes this case from trolley cases in which it's permissible to kill the one to save the five, together with a general principle to the effect that actions with that feature are always wrong. We must cite some such principle; otherwise the explanation is incomplete. And if the principle we cite admits of counterexamples, it's back to the drawing board; we haven't found the facts in virtue of which the act is wrong. (2017b, p.138)

Rosen talks of principles instead of models, but the effect is the same – when our models' predictions fail to be as we'd like them upon input of new descriptive facts, we'd better alter our model or come up with a new one altogether. We have shown, in philosophy, a tremendous ability to generate descriptive facts to input in the form of trolley cases.[11] And have revised our models for when one may turn the trolley in light of these cases, by positing new parameters for our models so they output predictions consistent with what we hope they'd output.

Of course I'm nowhere near the first to point out that trolleyology may be methodologically suspect for its reliance on (often recherché) cases (see, among many others, Fried, 2012,

[11] When one looks in Kamm's (2022) index, one finds 25 individual cases listed under the 'trolley problem' entry, and many of these cases have variations and sub-variations as well.

Bauman et al., 2014). What I'm doing is offering a diagnosis: this practice will seem suspect if we're worried — as we should be — about overfitting, as the more input-output pairs we fit our models to, the more we run the risk of overfitting.

The worry is not unique to trolley cases, either. In the moral responsibility literature, we will find a proliferation of Frankfurt cases, in which agents seem to act freely but could not have acted otherwise, given to put pressure on the reasonable-seeming principle that people are only morally responsible for what they do if they could have acted otherwise. These cases often feature "strange and esoteric" (Kane, 2007 p.168) science-fiction scenarios and abundant backup mechanisms, and we revise our theories of moral responsibility in light of them (see e.g., Sartorio, 2016). Cases also abound, and get pretty complex, in the self-defense literature (see e.g., Frowe and Parry, 2022 for a sampling), and theories of self-defense are revised in light of them in the same manner.

The same could be said for many more particular domains of ethical inquiry. This is standard – and, I think, worrying – methodology in ethics. In principle, it might turn out that we are making all the right choices in regard to the parameters to posit and the data to fit to when building models for trolleyology, moral responsibility, self-defense, and elsewhere. But to assert this seems optimistic, to say the least. It seems far more likely that we've made mistakes *somewhere*, as we proliferate increasingly complicated cases and alter models to capture them, positing potentially ad-hoc parameters that may give us bad results. And, at the very least, there is no commonly accepted methodological practice in ethics that we employ explicitly to guard against overfitting. This seems risky at best. So, in the following section, I put forward preliminary suggestions for such a practice.

4 How to Avoid Overfitting

The main task of this paper is to argue that some standard methodology in normative ethics has us running overfitting risks. Its task is not to give conditions or instructions for building non-overfitted models. But it would be remiss if it said nothing at all about this.

What, then, can be done to guard against overfitting? Not all solutions from the sciences are appropriate here. Scientists might like to remove an independent holdout sample of their data and use it to check a model post-construction (Hawkins, 2004), but it isn't obvious we should do this in ethics – how shall we pick the data to hold back? Do we have enough

good data that it's appropriate to decline to use some amount of it? These things aren't clear. Scientists might like to check whether their potentially overfitted model is more complex relative to another equally good model (Hawkins, 2004). Fine – we may reasonably take simplicity to be a tiebreaker in moral modeling as well. But we may take it merely as a tiebreaker. A simpler model must only be preferable to a more complex one if they are identical in all other relevant respects. And part of the problem is that it's difficult, in the moral realm, to know what respects are relevant. A more complex model might end up being more explanatory, or might capture a larger range of cases despite some inaccuracies, and it is not immediately obvious how these concerns weigh against complexity. Scientists might like to check for agreement among models (Salman and Liu, 2019), with ones that break from consensus being likelier candidates for having been overfitted. This isn't likely to help much in the current state of ethics. Our best moral models will generally agree on the easy cases like promise-breaking, and might predict different moral facts obtain in cases of, say, harming one to save many. It is this very disagreement that we want to figure out, and an argument to the effect of "more models tell us to harm one to save many, so that's what non-overfitted models will do" seems just obviously mistaken.

In fact, it's hard to diagnose overfitting by looking at models and their outputs in ethics in large part because it's often hard to tell when a moral model isn't functioning well. If I've built a model to predict the weather, I can see that I've done a bad job when it predicts clouds but the sky is clear. But if I've built a model to predict when one should turn the trolley, I don't necessarily know how to tell whether I've done a bad job. It's not so obvious when one should turn the trolley – certainly not as obvious as when it's cloudy outside. The point is: it's sometimes difficult to tell when moral models are giving us junk predictions. So this isn't a particularly useful way to tell whether they're overfitted.

Instead, we should look more closely at the data to which we're fitting our models. Overfitting occurs when a model is overly flexible by virtue of accommodating all data at the cost of other virtues such as predictiveness. So we had better not introduce ad-hoc components to accommodate all data at the cost of other virtues.

This will, in our state of moral and epistemic uncertainty, result in some data not being accommodated. This is a common solution to overfitting (see e.g., the literature on overfitting and Gettier cases – Weatherson, 2003, Weinberg, 2017, Williamson, 2021, 2024). To that end, we should be looking for decision procedures for a methodologically principled way to pick

which data to decline to accommodate. These would be requirements on data, not on models – but if these requirements are good, they should help us to get better models.

An obvious – but, I think, not terribly promising – requirement on data will be a requirement on the outputs in the input-output pairs to which we fit models. Again, these outputs capture the intuitive judgments about cases we try to align our models with. Maybe we should only take into account – and fit our models to – the outputs that are, well, correct. Of course the problem is that this advice amounts to a directive to just get better at ethics, or at having intuitions. Again, we can confidently say that Kant was wrong if he thought women could permissibly be subject to unequal marriages, but we can also bet that future philosophers will be pointing out obvious errors in our own judgments. Presently, there is an enormous literature — and a great deal of disagreement – on the uses and dangers of such intuitions in ethics, and on how we may tell the good from the bad (see McMahan, 2000, Hanson, 2002, Singer, 2005, Herok, 2023 for just a small sample). If any views in this vicinity turn out to be correct, then our problems will be solved. Until then, though, a directive to accommodate only correct outputs will be of little help.

We might throw up our hands and say we should simply decline to fit our models to any such outputs, if the whole business of intuition-mongering is so suspect (see Cappelen, 2012). This seems extreme – if we cannot take into account intuitions, then unless we have a different story to tell about how to try to check models' outputs directly against the actual moral facts, it is not clear how we should proceed. Most normative ethicists rely on intuitions to some degree, and want to continue doing so. I don't want to suggest such a radical departure from their practice.

Instead, then, let's take a closer look at the inputs in those input-output pairs. A requirement on such inputs seems more promising than a requirement on outputs. Again, these are the descriptive facts we feed into models to output predictions about the moral facts – the details of Trolley, for example. As noted, part of the worry when it comes to trolley cases, Frankfurt cases, and others is that some inputs seem intuitively appropriate to revise models in light of, and others seem less so. We should want some requirements to distinguish between these types of inputs. What follows, then, are some suggestions for the kinds of requirements that should do us good.

First, a quick detour to motivate such requirements: Weatherson (2003), Weinberg (2017), and Williamson (2024) argue that there are at least possible overfitting-related methodological

problems in the literature on Gettier cases. We cannot have full confidence in and therefore should not overfit our theories to this data, so it is at least conceivable that JTB-style theories of knowledge are not so threatened by Gettier as we might have thought. Maybe Gettier cases are just not the right kind of inputs to fit a theory to.

Williamson goes on to attempt to vindicate Gettier on non-methodological grounds (2024, see also 2000, 2013, 2015). But I want to make an argument for the usefulness of Gettier cases on methodological grounds – so, two observations about Gettier cases.

First, Gettier cases seem to share certain descriptive features. When I teach Gettier to undergraduates, put them into small groups, and give them five minutes to come up with their own Gettier cases to present, they almost always come back with perfectly good cases, indicating that the cases share some common characteristics. We might argue about what exactly those characteristics are – Zagzebski (1994) claims that we may generate Gettier cases by a rule involving picking some piece of bad luck that would, in ordinary circumstances, make an otherwise fine-seeming belief false, and then adding some separate piece of good luck such that the belief is true after all – but it seems like there's something there.

Second, Gettier cases also receive the same kinds of evaluations. This is not to claim that individuals in fact take Gettier cases to be instances of knowledge or not, as that is a matter of ongoing empirical dispute (see, e.g., Weinberg et al., 2001). But we're interested in whether all or any such cases are instances of knowledge, and would generally use the same tools for all cases to figure this out. It would be strange to think that Smith doesn't know that the man who will get the job has ten coins in his pocket but does know that either Jones owns a Ford or Brown is in Boston, and even stranger to think that the tools we'd use to figure out the former would be different from the tools we'd use to figure out the latter.

To the first point, surely it is good to have a simple, clean rule for generating data to input in a model. In this respect, the data generated by this rule is data in which we can be confident. Consider, by contrast, a set of inputs that we can only generate via a less clean rule – i.e., 'here's a bunch of inputs – we generate most of them via Rule A, then this small subset via Rule B, so our rule for generating such inputs is the disjunction of Rule A and Rule B.' So we should be happier about fitting to such inputs in our modeling than inputs generated by a more complex rule, or no rule at all. And to the second point, surely it is good to have a model take inputs that receive the same kind of evaluation. We shouldn't expect our model for predicting X by means M to deal well with cases where the thing to be predicted is Y by

means N.

So, on overfitting grounds, we should prioritize revising models in response to cases that we can generate via a simple rule giving a set of descriptive facts which all receive the same kind of evaluation. In the moral case, we should be thinking about descriptive facts that receive the same kind of moral evaluation specifically, since that's the domain we're interested in.

Here is how these concerns might apply in the practice of moral modeling. We can generate a case like Trolley via a simple rule about, say, situaitons in which we have the choice to harm one to save many. And with this rule we can generate other canonical cases:

> Bridge: You are on a footbridge watching a runaway trolley head towards five people stuck on the tracks. There is a man next to you on the bridge, and if you push him off the bridge onto the tracks, the trolley will hit and kill him, but not hit and kill the five. (Thomson, 1976)

So if our model is designed to predict moral facts when given Trolley as input, we should feel – on these grounds – pretty confident about using it for Bridge, and revising it if it goes clearly wrong.

Then there are some inputs that perhaps we should feel less confident about on these grounds. I am thinking here of the proliferation of baroque trolley cases in the literature. Take Kamm's Tractor Case, in which "the five toward whom the trolley is headed also have a deadly tractor headed toward them. If we turn the trolley away it will hit one person whose being hit will stop the tractor" (2015, p.68) and Tractor Case II, "which is like the original Tractor Case except that the person's being hit on the side track has no causal role in stopping the tractor. Rather, the tractor is stopped by a switch that is pressed by the trolley as it is turned away from the five" (2015, p.70). It is not clear to me that our original harm-one-to-save-many rule accounts for these two cases, along with both Bridge and Trolley, even if they're all to be evaluated morally. We might have to add some tractor parameter or no-causal-role parameter to our rule to generate these cases.

If we're searching for the moral law, we'd better take into account Tractor Cases as inputs – the moral law is, of course, exceptionless and universal, so should output the actual moral facts in all cases, including these. But moral models aren't subject to the same requirements. Maybe I need not take Tractor Cases into account when building a model that predicts what to do when one is in a position to harm one to save many. If such a model gives a bad-seeming

result when I input a Tractor Case, perhaps I should conclude that it was a mistake to give it a Tractor Case rather than revise it to accommodate Tractor Cases.

My point is this. We should want to fit models to our best, cleanest, and in some sense most reproducible inputs. This is a way to be careful and principled about our data, and thereby to avoid overfitting. And this meshes well with our motivations for embracing a modeling approach in the first place. As noted, we might have simplified, false, or mutually inconsistent models doing good work for us, and some of models may treat some particular inputs – and admit of revision based on these inputs – differently from other models.

Now, I do not mean to claim that I have given a fully general requirement on data here. This is a recommendation for theorizing rather than a set of necessary and sufficient conditions for the kinds of data we should deal with in a model-based practice of normative ethics. I do not give guidelines for how exactly we might tell a rule used to generate inputs is too complex, or exactly how we should identify our cleanest data. In fact, some inputs that do not fit these recommendations might end up being incredibly important on some other grounds – some cases relevantly like the Tractor Cases may be very important, and some relevantly like Bridge may be entirely unimportant, for reasons other than overfitting. But this is a common kind of conclusion in discussions about methodology. If some feature of a model is all-else-equal desirable, we still need to figure out whether all else really is equal. We'll have to do this work on a case-by-case-basis.

What I'm doing here, instead, is making a broad methodological claim: if we care about overfitting in our moral models, we should evaluate not only the models themselves but also the data we input with overfitting worries in mind. I am adding more methodological tools to draw upon when building moral models. So, in a way, what I am doing here is assigning ethicists more work. But this work, I think, could help us improve our practice, at the very least making us more aware of overfitting risks and giving us some strategies to address them.

5 Conclusion

I have presented reasons for thinking that normative ethical inquiry ought to – and does – make use of modeling, and identified overfitting as a peril of the model-building approach. If, as I argue, ethicists have been remiss in guarding against overfitting, then the suggestions

that I offer here will improve our practice. Insofar as we conceive of ethical inquiry as model-building – as an end to guide action, or as a means to get us closer to discovering moral laws – then paying closer attention to the data we feed into our models will help us be more methodologically principled and avoid bad results.

As noted, these worries apply particularly to the practice of normative ethics because of our counterexample-heavy methodology and our concern for correctly predicting moral facts. I certainly don't mean to say these worries apply *only* to ethics – they will apply to any philosophical subfield insofar as we care about counterexamples and predictiveness there, which is to say they will apply, to varying degrees, to all philosophical subfields. But, first: it is not widely recognized that ethicists, even those who do not use formal tools or methods, may productively engage – and are in fact engaging – in modeling practices, and may benefit from the use of scientific tools. And, second: it is also that I am an ethicist, and so it particularly matters to me that we get things right in this area. I hope this is one way that we can do so.

Acknowledgments

I'm grateful to Andy Egan, Alex Guerrero, Adrian Liu, Austen McDougal, Carolina Sartorio, Jordan Scott, Bram Vaassen, participants at the 2024 Chapel Hill Normativity Workshop, and anonymous referees for their feedback on (various drafts of) this paper.

References

Jonathan Alexander and Jonathan Weinberg. The 'Unreliability' of Epistemic Intuitions. In Edouard Machery and Elizabeth O'Neill, editors, *Current Controversies in Experimental Philosophy*. Routledge, 2014.

D. M. Armstrong. *What is a Law of Nature?* Cambridge University Press, 1983.

Christopher W. Bauman, A. Peter McGraw, Daniel M. Bartels, and Caleb Warren. Revisiting External Validity: Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology. *Social and Personality Psychology Compass*, 8(9):536–554, 2014.

Selim Berker. The Explanatory Ambitions of Moral Principles. *Noûs*, 53(4):904–936, 2018.

Herman Cappelen. *Philosophy without Intuitions*. Oxford University Press, 2012.

Nancy Cartwright. *How the Laws of Physics Lie*. Oxford University Press, 1983.

David Copp. *Morality in a Natural World: Selected Essays in Metaethics*. Cambridge Studies in Philosophy. Cambridge University Press, 2007.

Jonathan Dancy. *Ethics Without Principles*. Oxford University Press, 2004.

Stephen Darwall, Allan Gibbard, and Peter Railton. Toward Fin de siècle Ethics: Some Trends. *The Philosophical Review*, 101(1):115–189, 1992.

Billy Dunaway. Supervenience Arguments and Normative Non-Naturalism. *Philosophy and Phenomenological Research*, 91(3):627–655, 2014.

Ronald Dworkin. *Taking Rights Seriously*. Harvard University Press, Cambridge, Mass, 1978.

Nina Emery. The Governing Conception of Laws. *Ergo*, 9, 2022.

Nina Emery. *Naturalism Beyond the Limits of Science: How Scientific Methodology Can and Should Shape Philosophical Theorizing*. Oxford University Press, 2023.

David Enoch. How Principles Ground. *Oxford Studies in Metaethics*, 14:1–22, 2019.

Daniel Fogal and Olle Risberg. The Metaphysics of Moral Explanations. pages 170–194, 2020.

Malcolm Forster and Elliott Sober. How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35, 1994.

Barbara H. Fried. What Does Matter? The Case for Killing the Trolley Problem (Or Letting It Die). *The Philosophical Quarterly*, 62(248):505–529, 2012.

Helen Frowe and Jonathan Parry. Self-Defense. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2022.

Peter Godfrey-Smith. The strategy of model-based science. *Biology and Philosophy*, 21(5): 725–740, November 2006.

Bart Gruzalski. Foreseeable consequence utilitarianism. *Australasian Journal of Philosophy*, 59 (2):163–176, 1981.

Andrew Hamilton. Laws of Biology, Laws of Nature: Problems and (Dis)Solutions. *Philosophy Compass*, 2(3):592–610, 2007.

Robin Hanson. Why Health is Not Special: Errors in Evolved Bioethics Intuitions. *Social Philosophy and Policy*, 19(2):153–179, 2002.

Douglas M. Hawkins. The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, 44(1):1–12, 2004.

Tomasz Herok. Intuitions are never used as evidence in ethics. *Synthese*, 201(2):42, 2023.

Michael Townsen Hicks. Dynamic Humeanism. *British Journal for the Philosophy of Science*, 69 (4):983–1007, 2017.

Christopher Hitchcock and Elliott Sober. Prediction Versus Accommodation and the Risk of Overfitting. *The British Journal for the Philosophy of Science*, 55(1):1–34, 2004.

Colin Howson. Fitting your theory to the facts: Probably not such a bad thing after all. *Minnesota Studies in the Philosophy of Science*, 14:224–44, 1990.

F. M. Kamm. *Rights and Their Limits: In Theory, Cases, and Pandemics*. Oxford University Press, 2022.

F.M. Kamm. *The Trolley Problem Mysteries*. Oxford University Press, 2015.

Robert Kane. Response to Fischer, Pereboom, and Vargas. In John Martin Fischer, Robert Kane, Derk Pereboom, and Manuel Vargas, editors, *Four Views on Free Will*. Wiley-Blackwell, 2007.

Immanuel Kant. *Practical Philosophy*. The Cambridge Edition of the Works of Immanuel Kant. Cambridge University Press, 1996.

Thomas Kelly and Sarah McGrath. Is Reflective Equilibrium Enough? *Philosophical Perspectives*, 24(1):325–359, 2010.

Pauline Kleingeld. On Dealing with Kant's Sexism and Racism. *SGIR Review*, 2(2):3–22, 2019.

Rae Langton. Duty and Desolation. *Philosophy*, 67(262):481–505, 1992.

Stephanie Leary. Non-Naturalism and Normative Necessities. *Oxford Studies in Metaethics*, 2017.

David Lewis. *Counterfactuals*. Blackwell, 1973.

Christian Loew and Siegfried Jaag. Making Best Systems Best for Us. *Synthese*, 197(6): 2525–2550, 2018.

Barry Loewer. Two Accounts of Laws and Time. *Philosophical Studies*, 160(1):115–137, 2012.

Jeff McMahan. Moral Intuition. In Hugh LaFollette, editor, *The Blackwell Guide to Ethical Theory*, pages 92–110. Blackwell, 2000.

Tristram McPherson. The Methodological Irrelevance of Reflective Equilibrium. In Chris Daly, editor, *The Palgrave Handbook of Philosophical Methods*, pages 652–674. Palgrave Macmillan, 2015.

Tristram McPherson. Supervenience in Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. 2019.

Mary S. Morgan and Margaret Morrison. *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press, 1999.

L. A. Paul. Metaphysics as modeling: the handmaiden's tale. *Philosophical Studies*, 160(1): 1–29, August 2012.

Karl Popper. *The Logic of Scientific Discovery*. Hutchinson, 1959.

Karl Popper. *Conjectures and Refutations: The Growth of Scientific Knowledge*. Basic Books, 1962.

Angela Potochnik. *Idealization and the Aims of Science*. University of Chicago Press, 2020.

Hilary Putnam. *Mathematics, Matter and Method*. Cambridge University Press, 1979.

W. V. Quine. The Scope and Language of Science. *The British Journal for the Philosophy of Science*, 8(29):1–17, 1957.

Gideon Rosen. Ground by Law. *Philosophical Issues*, 27(1):279–301, 2017a.

Gideon Rosen. What is a Moral Law? *Oxford Studies in Metaethics*, 12, 2017b.

Bertrand Russell. *The Problems of Philosophy*. William & Norgate, 1912.

Shaeke Salman and Xiuwen Liu. Overfitting Mechanism and Avoidance in Deep Neural Networks, January 2019.

Carolina Sartorio. Frankfurt-Style Examples. In Kevin Timpe, Meghan Griffith, and Neil Levy, editors, *The Routledge Companion to Free Will*. Routledge, 2016.

Russ Shafer-Landau. *Moral Realism: A Defence*. Oxford University Press, 2003.

Marcus G. Singer. Actual Consequence Utilitarianism. *Mind*, 86(341):67–77, 1977.

Peter Singer. *Animal Liberation*. HarperCollins, 1975.

Peter Singer. Ethics and Intuitions. *The Journal of Ethics*, 9(3/4):331–352, 2005.

S. J. Singer and Garth L. Nicolson. The Fluid Mosaic Model of the Structure of Cell Membranes. *Science*, 175(4023):720–731, February 1972.

Michael Smith. *The Moral Problem*. Blackwell, 1994.

Bart Streumer. *Unbelievable Errors: An Error Theory About All Normative Judgments*. Oxford University Press, 2017.

Bart Streumer. Standing up for supervenience. *Philosophy and Phenomenological Research*, 109 (1):138–154, 2024.

Judith Jarvis Thomson. Killing, Letting Die, and the Trolley Problem. *The Monist*, 59(2): 204–217, 1976.

Judith Jarvis Thomson. The Trolley Problem. *The Yale Law Journal*, 94(6):1395–1415, 1985.

Peter Unger. *Living High and Letting Die: Our Illusion of Innocence*. Oxford University Press, 1996.

Brian Weatherson. What Good Are Counterexamples? *Philosophical Studies*, 115(1):1–31, 2003.

Jonathan Weinberg. Knowledge, Noise, and Curve-Fitting: A Methodological Argument for Justified True Belief? In Rodrigo Borges, Claudio de Almeida, and Peter D. Klein, editors, *Explaining Knowledge: New Essays on the Gettier Problem*, pages 253–272. Oxford University Press, 2017.

Jonathan M. Weinberg, Shaun Nichols, and Stephen Stich. Normativity and Epistemic Intuitions. *Philosophical Topics,*, 29(1-2):429–460, 2001.

Michael Weisberg. *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press, 2013.

Michael Weisberg. Modeling. In Herman Cappelen, Tamar Szabó Gendler, and John Hawthorne, editors, *The Oxford Handbook of Philosophical Methodology*. Oxford University Press, 2016.

Timothy Williamson. *Knowledge and Its Limits*. Oxford University Press, 2000.

Timothy Williamson. Gettier Cases in Epistemic Logic. *Inquiry*, 56(1):1–14, 2013.

Timothy Williamson. A note on Gettier cases in epistemic logic. *Philosophical Studies*, 172(1): 129–140, 2015.

Timothy Williamson. Conclusion: Semantics, Heuristics, Pragmatics. In *Suppose and Tell: The Semantics and Heuristics of Conditionals*, pages 264–266. Oxford University Press, 2020.

Timothy Williamson. Degrees of Freedom: Is Good Philosophy Bad Science? *Disputatio*, 13 (61):73–94, 2021.

Timothy Williamson. *Overfitting and Heuristics in Philosophy*. The Rutgers Lectures in Philosophy. Oxford University Press, 2024.

Tobias Wilsch. The Governance of Laws of Nature: Guidance and Production. *Philosophical Studies*, 178(3):909–933, 2020.

William Wimsatt. False Models as Means to Truer Theories. In Matthew Nitecki and Antoni Hoffman, editors, *Neutral Models in Biology*, pages 23–55. Oxford University Press, 1987.

Xue Ying. An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, 1168(2), 2019.

Linda Zagzebski. The Inescapability of Gettier Problems. *The Philosophical Quarterly*, 44(174): 65–73, 1994.