Ethical Theory and Moral Practice
https://doi.org/10.1007/s10677-022-10326-4

Modelling in Normative Ethics

Joe Roussos[1]

Accepted: 16 September 2022

Abstract

This is a paper about the methodology of normative ethics. I claim that much work in normative ethics can be interpreted as modelling, the form of inquiry familiar from science, involving idealised representations. I begin with the anti-theory debate in ethics, and note that the debate utilises the vocabulary of scientific theories without recognising the role models play in science. I characterise modelling, and show that work with these characteristics is common in ethics. This establishing the plausibility of my interpretation. Taking methodological inspiration from modelling in science gives us new tools for managing idealisations, and a new perspective on pluralism. I think demonstrate why this interpretation is a fruitful way of interpreting ethics, by looking at three case studies. First, I return to the anti-theory debate and argue that modelling opens up a new middle ground. Second, I argue that a modelling lens offers a new way of understanding impossibility theorems in population ethics, and their bearing on ethics as a whole. Finally, I show how viewing our work as modelling can be deployed in debates within ethics, using the debate over prioritarianism as an example. I close with further methodological suggestions for those who choose to see themselves as modellers. I discuss the role of counterexamples, our responses to moral disagreement, and the training of new ethicists.

Keywords Moral theory · Anti-theory · Modelling · Methodology · Population ethics · Prioritarianism

1 Introduction

This is a paper about the methodology of normative ethics. It is exploratory and intended to be provocative. What I am exploring is the notion that moral philosophy is engaged in modelling.

What is a model? A favourite example will get us started, before the more detailed characterisation below. Imagine that I am studying the fish population in my local pond. I observe the fish feeding, breeding, and dying, for a few generations. I realise that the pond has a finite carrying capacity for fish, due to their needs for space and competition for food. I observe that the population this week depends positively on the population last

✉ Joe Roussos
joe.roussos@iffs.se

[1] Institute for Futures Studies, Stockholm, Sweden

Published online: 03 October 2022

week, but that as the population reaches the capacity of the pond, crowding hampers population growth. Reflecting on these patterns, I decide to use the following equation to predict changes in the fish population: $N_{t+1} = 4N_t(1 - N_t)$, where N is the number of fish in the pond divided by the carrying capacity, and t is a time index counting months.

In so doing, I am modelling the fish population. This involves representing the fish population, in my case mathematically. Only certain features of the real pond and fish are represented, however; I have ignored the natural variation in fish size and reproduction. I have also ignored factors which I know to influence the population level of the actual pond, such as fishing. I treat time as discrete, and count in months. I make no claims that this equation describes fish growth everywhere: the form of the equation is chosen to fit the rate of reproduction of this population. The features that I will take as characteristic of modelling in this paper are these: (1) I *represent* the fish pond, in this case, using mathematics; (2) this representation is *idealised*: it leaves out some properties and adds in others which the real pond lacks; and (3) the idealised representation acts as a *proxy*, I study it to learn about the population in the pond.

This paper fits into a growing discussion about modelling in philosophy. Williamson (2006, 2017, 2018) has argued repeatedly that modelling is an important method in philosophy. Various authors have debated whether regarding metaphysics as a modelling discipline might illuminate its methodology, and confer upon it some scientific respectability (Godfrey-Smith 2006, 2012; Paul 2012; Novick 2017). In formal epistemology it is increasingly common for philosophers to describe their own practice as modelling (e.g., Bovens and Hartmann 2003; Eva and Hartmann 2020), and to advocate for it as a method (Leitgeb 2013, 273). Recently, Beck and Jahn (2021) proposed success conditions for models aiming at normative guidance, and Roussos (forthcoming) argued that normative formal epistemology must be modelling, as other interpretations of its methodology make it a failed exercise.

In all these cases, the talk of modelling and model-building is an analogy with the common-place scientific practice of inquiry using idealised representations, which we have just met. I share with these writers a common goal: to illuminate and perhaps improve philosophical methodology by thinking about existing commonalities with scientific methodology, and to advocate for the adoption of new methods. Here I shall do so for one part of moral philosophy. I take moral philosophy to be a broad heading which covers metaethics, descriptive or empirical ethics, deontic logic, and my focus here: first-order normative ethics. By this I mean the study of goodness and of right action which aims to provide norms and principles that govern moral behaviour. My claim is that normative ethics can and does make use of idealised models like my model of the fish population in my pond. I also offer some reason to think that this is a good thing, though this paper does not aim to encourage ethicists to start modelling.

The claim that ethics makes use of modelling today is an interpretative claim about methodology. The empirical claim there is *some* modelling in ethics is trivial: there are clear instances of conscious, self-described modelling in formal ethics, typically in the welfarist tradition. My discussion below will vindicate this self-description. But my claim here is broader. Much work in ethics shares characteristic features with modelling in science, be it formal or informal, consequentialist or non-consequentialist. The similarities between this work and the explicit modelling in ethics has been masked by superficial dissimilarities, such as the use of mathematics. I claim that this latter, much larger, part of ethics can fruitfully be understood as modelling.

In large part, the ethicists I describe do not currently see themselves as modellers. But this shift in perspective has a number of benefits. Modelling is a well-established method in science, with four or five decades of methodological study in philosophy of science dedicated to it. Ethical modelling can benefit from adopting the methodological lessons about good modelling from this literature. Of particular importance are tools for carefully managing the impact of idealisation. The ineliminable nature of idealisation in models leads to a kind of model pluralism, which involves accepting that certain apparent disagreements cannot be resolved. It also matters for how we argue: models are not sensitive to counterexamples in the way that much of ethical theory is taken to be. So, if ethicists come to see themselves as modellers, this may bring about a significant shift in how ethicists practice their craft.[1]

I also demonstrate several benefits to understanding (bits of) ethics as modelling, with reference to current debates. The first is a new way of understanding what is going on (and going wrong) in the theory/anti-theory debate in ethics. The second is a new way of understanding impossibility theorems in population ethics, and their bearing on ethics as a whole. Third, I show how the fact that we are modelling can be deployed within debates in ethics, using the debate over prioritarianism as an example.

2 Theory versus Anti-Theory

As a methodologist, I like to take as my starting point a disagreement between practitioners. So let us reflect on a bitter conflict in normative ethics: the "anti-theory" debate. Anti-theorists have long criticised a certain kind of ethical philosophising as misguided, doomed to fail, and besides the point. The target of their critique is given the name "theory", sometimes "ethical theory" or "moral theory". So what is a moral theory?

While the term is much contested, here are some characteristics attested to by both theorists and anti-theorists. A moral theory:

- provides a decision procedure for determining which actions are right or wrong (Chappell 2022; Fotion 2014; Louden 1992; Nussbaum 2000; Timmons 2012; Williams 1981)
- is general, or complete, or universal; i.e., it applies to all of, or a very wide range of, circumstances, people, action-types, and so forth (Chappell 2022; Louden 1990; Fotion 2014; Nussbaum 2000)
- is or aspires to be the uniquely true theory. Theories compete; there is only one correct (complete) theory (Chappell 2022; Fotion 2014).
- is decidable; i.e., one can check whether any particular action or belief is correct according to the theory (Fotion 2014; Nussbaum 2000)
- is axiomatisable; i.e., can be stated by theorists in terms of a finite set of principles (Fotion 2014; Louden 1992)[2]

[1] Note that I am talking here of seeking an interpretation which best fits with current practice. I seek to rationalise current methodology, rather than to describe what ethicists take themselves to be doing. Thus I avoid speaking about whether ethicists are "really" modelling—a question which likely turns on psychological facts I am not interested in.

[2] Though I will focus on normative ethics, this conception of theory is taken up more broadly in moral philosophy. List and Valentini (2016, 15–16), writing about political theory, use a definition of "theory" which includes a list very similar to the above.

Some combination of the above is often taken to mean that there are no moral dilemmas (Louden 1992, 97); given a complete description of the circumstances, the moral theory yields a single consistent verdict.[3] In addition to ruling actions right or wrong, moral theory is supposed to tell us what *makes* these actions right, thereby offering an explanation of their rightness (the same goes, mutatis mutandi, for goodmakers) (Timmons 2012, 13).

Anti-theorists often take utilitarianism as the paradigm of a (problematic) theory; in its more ambitious forms, it exhibits all of these characteristics.[4] The theory of rights and Kant's ethics are two other prominent examples and recipients of anti-theory criticism (O'Neill 1987). I will note immediately that, in practice, the term "theory" is applied much more widely than the list above might suggest. Work which aims at less than universal scope, or which is open to some pluralism, or which has unclear decidability and axiomatisability, is nevertheless called ethical theory. I suspect that the core notion has to do with systematicity and wide scope—i.e., the first two bullets on the list—with uniqueness next-most central. I won't try to pin down exactly what ethicists mean by theory, since one thing I want to suggest in this paper is that some "theorising" instead be understood as "modelling".[5]

It is helpful to distinguish here between different potential objects of methodological analysis. First, there are moral theories: Kantianism, say, under its most imperial interpretation. Then, there is the project of moral theory: the pursuit of a theory with maximal scope and generality, meeting all the criteria above. Finally, there are pieces of philosophical work, as presented in articles and papers. The theoretical constructs involved in these pieces of work are what I will mostly be interested in, and they often fall into the category of things described as theory but which don't have all of the features on the list above. Ethicists often describe them as though they are natural extensions of a theory: "what Kant's ethics says about childcare," or whatever. Many of these pieces of work, I will argue, can be understood as models. I will also consider, in Sect. 4, a more radical re-interpretation in which self-described moral theories like Kantianism are models.

Returning to the list: this notion of "theory" is similar to and draws on the concept of a scientific theory. Scientific theories are tightly associated with laws of nature, regularities that are taken to hold very generally in a domain. Science begins with observations of particular facts, and proceeds by noticing certain patterns in the empirical phenomena. These patterns, sometimes called empirical laws, are one thing that science seeks to explain through theories. Theories aim to unify diverse phenomena by presenting the empirical uniformity they exhibit as the results of a common set of basic theoretical laws (Hempel 1966, 75). Theories seek to explain that uniformity, offering explanations and understanding of the phenomena in question, and allowing for predictions via the laws. Theoretical

[3] Occasionally it is also assumed that the theory is a set of sentences, which is consistent and deductively closed. This ensures that "theory" has the same meaning in ethics as it has in mathematical logic, and depending on how one formalises things, may make the no-moral-dilemmas aspect a consequence of this definition. But this is not essential; there are consistent axiomatic theories in deontic logic in which there are moral dilemmas.

[4] E.g., Anscombe (1958) blames Sidgwick for bringing about the negative change that she detects in all English-language moral philosophy after him. Williams is another clear case, notably in (Williams 1973, 1981).

[5] So, for example, I am not concerned that O'Neill (1987, 59) rejects the "decision procedure" and "universality" conditions as reasonable requirements for what she calls ethical theory. I discuss narrow scope theorising with purposes other than decision determination below, as a kind of modelling. This kind of terminological confusion is inevitable in a project like mine, which reflects on current methodology and suggests new methodological categories.

laws involve the introduction of theoretical concepts, which go beyond what can directly be observed.

In the late nineteenth century, philosophers began to analyse mathematical and scientific theories in formal languages (see Glymour 1999 for a historical discussion). In one resulting tradition, theories came to be understood as sets of sentences in such a formal language. These sets are consistent, deductively closed, and (ideally) axiomatisable. It is from this tradition that we get the logical terminology used to characterise moral theories in the list above.[6]

Why do "anti-theorists" object to theory, so described? There is an entire literature of arguments on this topic, so I will note but a few. The common thread between the arguments I have selected here is that the anti-theorists take themselves to be interested in *the way things really are*. They focus on the texture of moral life, or on applications to real moral problems, or on the actual process of moral deliberation.

First, theory simplifies too much; it removes the nuance, complexity, and difficulty of moral reasoning. Bernard Williams is famous for this critique of utilitarianism. Reflecting on a pair of cases that he takes to be dilemmas, but which utilitarianism has ready answers for, he writes: "Not only does utilitarianism give these answers but, if the situations are essentially as described and there are no other special factors, it regards them, it seems to me, as *obviously* the right answers. But many of us would certainly wonder whether…that could possibly be the right answer at all; and…even one who came to think that perhaps that was the answer, might well wonder whether it was obviously the answer" (Williams 1973, 99). The point is not that utilitarianism arrives at the wrong answer, but that it oversimplifies. This critique is not restricted to utilitarianism, either; McKeever and Ridge (2015) cite Raphael (1974) as deploying the same argumentative strategy against Kantianism.

Relatedly, theory is said to be too abstract, and too coarse, to deliver usefully precise recommendations in real situations. Arras (2010) claims that theory will often leave too many options on the table, not because they are in truth morally equivalent but because theories are incomplete in an important sense. Any theory will "run out of gas before it reaches the level of concrete decision making required by practical ethics" (Arras 2010, S3.3).

Next, there are too many moral theories. For Baier (1989), the very proliferation of theories shows that they are unlikely to succeed. This is because the ethical domain is simply too diverse to support successful theorising; it is not unified in the way that is required for theory to succeed. "Where do we have genuine and useful theories? Primarily in the sciences—but there we find a plurality of them primarily over time, rather than at a time. We certainly do not find some engineers building bridges or spaceships by application of one theory, while others at the same time are applying another different theory" (Baier 1989, 33–34). For Arras (2010, S3.1), this glut of contrary theories without an obvious choice between them makes theory useless for practical ethics, in the sense of providing a useful guide to practice.

Finally, theories require "principles which are definite in meaning in order for them to play their role in the deduction of particular moral judgements. On the other hand, the norms of actual moral practices are vague in order to permit context to play a role

[6] I am not claiming that the syntactic conception of theory is the *introduction* of the requirements of axiomatisation etc. Indeed, Aristotle's philosophy of science has a central role for axioms, deduction, and consistency. I am here merely highlighting the *identification* of "theory" with such a set of sentences.

in determining their application" (Clarke 1987, 238). As Baier argues, a seemingly clear norm such as "don't kill" "brings with it a very rich cultural baggage, if it is to have any content at all. Either it is a purely formal moral code, not yet prohibiting or enjoining anything, or else the form gets its determinate filling, in which case we are committed not merely to these 'negative' rules but to the rules of background institutions and ways of life that supply the determinate content to these prohibitions." Theories, with their focus on the norms alone, are thus unable to stand in the required justificatory relation to actual moral practices (Baier 1985, 273–74, quoted in Clarke).

There might be something right to these anti-theory critiques. But there is something deeply wrong with how the space of methodological options is characterised in this debate. Anti-theorists often seem to take the options to be "theory" (bad), or a form of very granular, piecemeal analysis that makes no attempt at generality or systematicity (good). In so doing, they neglect a middle-ground of partial systematisation, making use of intermediate principles with application to limited but still substantial domains. In science, models cover this middle-ground.

3 Interpreting Ethics as Modelling

In this section I present an interpretation of ethics under which much of it is modelling. I do so by describing characteristic features of modelling in science, and showing that the ethical work I am interested in has these features. This secures the plausibility of the modelling view: not only does the work have these features, but the modelling interpretation makes sense of some practices in ethics—why ethicists talk in certain ways, and why their work functions as it does.

I will now describe what it is to be a model, in terms of characteristic features.[7] There are numerous kinds of models in science, and I will describe various kinds in ethics too. What they share are the characteristics described below, which come with some methodological norms which I claim ethics should also adopt.

Here are the characteristic features: Models are idealised representations, which form part of an indirect strategy of inquiry (called modelling).[8] I will go through these features, explaining how they work in science and in what sense ethical work shares them.

3.1 Representation and Indirect Inquiry

Scientific models are representational in two senses, often called representation-of and representation-as. Many scientific models are representations *of* real systems, which are called the "target" of the model. These can be either specific systems like my pond or kinds

[7] I will not give a precise definition, or metaphysical account of what it means to be a model in normative ethics. This is a deliberate choice. Philosophers of science disagree over whether a single such account is possible for all scientific models, whether it is useful to provide one, and, if yes to both of these questions, which candidate account is correct. I think the answer to the first two questions is "no", and I think taking a stance on this philosophy of science dispute is unnecessary for my present purposes.

[8] My use of the term "model" differs from the use in formal semantics and logic, on which a model is a mathematical structure satisfying a set of sentences. Some philosophers of science (e.g., Suppes (1969)) have argued we should understand all models in science as being models in this sense, but I do not think this is a plausible analysis of *all* scientific models. My use of "model", explained in this section, is wider than the semantics use.

of system, like a predator–prey system. Without going too deeply into the theory of representation-of, we can note that it involves two systems of objects, one of which stands for or denotes the other. In the opening example, the mathematical variable *N* stood for the population density of the fish pond. Models also represent the world *as* being a certain way—typically a way which is simpler and different from how the world actually is. Some models don't have real systems as their targets, but they are nevertheless representations-as, just as a picture of a dragon is a kind of representation although there are no dragons.

Scientific inquiry with models is "indirect" in that the scientist spends their time working with and studying the model, as a proxy for the target system. Rather than counting fish in the pond, I manipulate the mathematical model and then make inferences about the fish pond.

Work in ethics is often representational. This is clearest in formal ethics, where mathematical objects explicitly represent people, or welfare, or side-constraints, or whatever. Conveniently, such work is performed by philosophers with mathematical and scientific training who themselves speak in terms of models and modelling. For example, McCarthy et al. (2019) present a social aggregation theorem. Mathematical structures are introduced which represent things in the domain being studied: populations, values, preferences. Ethical principles are likewise represented in the model, as mathematical constraints or relations or properties of objects. This work is also indirect inquiry: the purpose of studying the mathematical structure is to learn about something else, the relation between individual and social welfare.

A common use of models in formal ethics is to test the consistency of a set of claims, or put another way, the compatibility of a set of conditions. It is often difficult to do this work without a model, because it is hard to see what exactly ethical principles imply. The modeller creates a model system, typically a mathematical structure, and the ethical principles or claims are translated into precise formal statements about it. One can then test whether a set of conditions is mutually satisfiable, using the precision afforded by the mathematics. This is modelling, in that it is indirect inquiry which studies a proxy system—the mathematical structure—in order to learn about the target—say, the betterness relation. In such cases there is room to argue about the faithfulness of the representation: whether the mathematical properties accurately capture the ethical principles which are their targets.

The use of representation in ethics goes beyond heavily mathematical work. Another obvious way in which work in ethics can be representational is in the use of diagrams, such as Parfit's (1984) box diagrams of populations. A less obvious, but very common, form of representational work in ethics is in the use of exemplary cases. I use the term "exemplification" here to mean "being an instance of a class, which also represents the class". For example, a paint swatch is an instance of a particular paint colour, and also represents that colour more broadly. The commonplace use of vignettes or cases in ethics involves exemplification. These are particular, imagined, situations which are intended to represent a wider class of situations, including real situations.

3.2 Idealisation

The mere use of representation is not enough for modelling, however. Models are characteristically *idealised*. Scientists typically cannot represent the systems they study completely accurately, either because the systems are too complex, or because their understanding is too limited, or because such a faithful representation would be intractable for analysis. So, in building their models, scientists leave out certain aspects of the system which they take to be irrelevant, and they represent the system as having properties that are different from its actual properties. These changes are called “idealisations” (Weisberg 2007a; Frigg and Hartmann 2018). Note that this term has no moral valence in science. Models are not thought to represent ideal systems in the sense of perfect or good systems. They are merely idealised as in different from reality in a way which makes them easier to study.

Distorting idealisations represent the target system as having properties different from its actual properties. For example, an inclined plane might be represented as frictionless while in fact it has friction. Often, the justification for this move is pragmatic: it simplifies the analysis. A model with friction might be much more complicated, perhaps too complex to be tractable. The idealisation may also facilitate focus: perhaps this investigation is about the contribution of gravity to a ball’s motion down the inclined plane, and not about the contribution of friction. Leaving out idealisations, also called abstractions, are equally prevalent in science. Here, scientists strip away properties of the target system which are assumed to be irrelevant. For example, in a Newtonian representation of a planetary system, planets might be represented as point masses, with positions that are a function of time. All other properties of planets are neglected: their shape, volume, colour, etc.

The line between distorting and leaving out idealisations is blurry. The justification for an idealisation can fit both descriptions, for example if one excludes friction in a case when it is known to be low but to make a small difference. Whether an idealisation removes or adds in also depends on how one frames the relevant properties: removing friction might be seen as adding slipperiness.

Note an important difference between how I use the term “idealisation” and how it has been used in prior discussions of abstraction in ethics, in particular the discussion following O’Neill (1987). O’Neill (1987, 56–57) makes the distinction I make here, noting that “the objection [to abstract theorising] is not just that much (too much) that is true of human agents is *omitted* in some accounts of agents, but that much (too much) that is false of human agents is *added*”, and calls this second change “idealisation“. However, she immediately notes that these same theories “idealise” in a moral sense: “they also treat enhanced versions of certain capacities as *ideals* for human action.” So for her, “idealisation” means “distortion” *and* “moral ideal”. I use “idealisation” as an umbrella term, for two reasons: (1) the line between leaving out and adding in is blurry and perspectival, and (2) not all distorting idealisations carry this moral sense.[9]

Science makes frequent and seemingly ineliminable use of idealisation. But the presence of idealisations means that models contain known falsehoods. How do we square these facts? This is a complex topic with a large literature dedicated to it (for an excellent recent entry see Potochnik 2017). One recent thread emphasises that humans are inquirers with limited cognitive capacities, confronting a hugely complex reality. Idealisations enable us to manage that complexity, by isolating particular aspects of nature for study. When it works well, idealisation focuses attention on a real and important factor, sometimes by highlighting its salience to the researcher, sometimes by freeing it from its interactions with other factors. Clearly, not all idealisation is good and recent work has focused

[9] Note that, because of this, there is no direct connection between my use of “idealised” and “ideal theory” in political philosophy. Non-ideal theory can make use of idealised models, as Hancox-Li (2017) points out in a discussion of models of racism and sexism. Such models may contain elements that are normatively ideal, such as perfect rationality, but be used to study distinctly non-ideal behaviour, such as self-segregation.

on characterising when it works well. For our purposes, the important lesson from that literature is that the success conditions are relative to the purposes of the inquiry, the inquirer's capabilities, and the system being studied.

This highlights an important feature of models, and a difference between models and theories: models are purpose-specific tools of inquiry. The purposes of inquiry, inquirer's capabilities, and eventual idealisations together set a domain of application for the model—outside of which it should not be expected to work well (Teller 2001; Weisberg 2007b). This feature of modelling explains why we encounter multiple, disagreeing scientific models of the same phenomenon. Teller illustrates this with an example of two models of water. The first is interested in the flow of water and wave propagation, and it represents the liquid as a continuous incompressible medium. The second is interested in explaining diffusion, say of a drop of ink in water. It represents water as a collection of discrete particles in thermal motion. Each is similar to water in the respects that are relevant to its purpose, but the two models look very different (Teller 2001, 401). Each is highly successful at its purpose, i.e., prediction of the relevant kind of behaviour, and their respective idealisations work well within their domain. But clearly they contradict one another: one says that water has particles, the other says it does not. The lesson is that neither should be thought to provide a definite characterisation of water, and our understanding of water is enhanced by having both available.

Distortion and abstraction are widespread in ethics. O'Neill (1987, 55) notes that abstraction is unavoidable in the search for principles of wide scope: "only abstract principles are likely to have wide scope: if ethical principles are to be relevant to a wide range of situations or of agents, they surely not merely *may* but *must* be abstract." But ethicists leave out more than merely what is taken to be irrelevant. Perhaps the most widespread example is the use of ceteris paribus clauses which assume that "all other factors are equal". A variant of this is when the ethicist simply ignores a factor that they acknowledge to be morally relevant and which cannot be held equal, such as the non-identity problem. (The usual justification is that including it would make things significantly more complex—a common justification for idealisation in science. There, the hope is often that future scientists will revisit this assumption and account for the neglected effect, which is perhaps the hope of ethicists in the case of the non-identity problem.) Indeed, one can view all instances of setting aside some moral considerations in order to focus on others as a form of idealisation. When we propose to study, say, duties of reciprocity separately from all other moral considerations, we are performing our inquiry under distorted conditions, isolating an aspect of morality which in reality is embedded in and interacts with a much more complex environment.

There are numerous other forms of idealisation in ethics, and in my discussion of the case studies in Sect. 5 and 6 I look at a few in detail.

Idealisation alone does not make for modelling; the three characteristics come as a package. Idealisation within in a proxy system—a representation used for indirect inquiry—is what is characteristic of modelling. But modelling *is* the natural home of idealisation, since the separation between target system and model system is a methodological device which is useful for managing the potential detriments of idealisation. One constructs the model, using whichever abstractions and distortions are deemed useful. The model is then manipulated, studied, used to generate some conclusions. But these are conclusions *about* the model. In inferring about the target system, what was *modelled*, one must consider how the idealisations influenced those conclusions.

Idealisations often limit the domain in which a model's results can be expected to hold. If one assumes that "other things are held equal" then it is important that the result not be assumed to hold when other things are not equal. Moreover, some things cannot be held equal, and there is a significant difference between assuming a factor is unchanging and ignoring it entirely. Care needs to be taken when inferring from model results which depend on such assumptions. This is difficult enough in science, where the models are descriptive. But in ethics our models are normative, and we must determine how to infer from a normative result generated by an idealised model to a normative conclusion about the target bit of morality in the real world.

More subtly, models often contain artefacts: properties of the model system that are not representative of any real feature of the target system but instead emerge from the representational choices of the modeller or the idealisations in the model. Good modellers must identify artefacts and ensure that they aren't imputed to the target. A common method for identifying such effects is "sensitivity analysis". The modeller varies idealisations—introducing minor air resistance, considering non-spherical canon balls—to test whether the results of interest are robust or mere artefacts of the unrealistic assumptions made to simplify the analysis.

Not all idealisations can or should be removed, however—the analysis may simply be impossible without them. This is one reason why modellers in science often make use of multiple models each offering a perspective on the target system. In ethics, this might look like a collection of models of the same aspect of morality, which make inconsistent assumptions and disagree on certain issues. Each has a purpose which guided its construction, and it is relative to that purpose that its results should be evaluated and made use of.

These are ideas and practices which exist in ethics, but seeing ethical work as modelling can help to unify and systematise them.

3.3 Models and Theories

What then is the relation between a model and a theory? As I noted at the outset, science contains many types of models and one way in which they differ is in this relation (Frigg and Hartmann 2018). Some help connect theories with reality, by representing a system of interest and subjecting it to the laws contained in that theory. Here is a high-school physics example: Newton's theory of mechanics consists roughly of his three laws along with some core notions like "centre of mass" and "reference frame", and a value for the parameter representing Earth's gravitational attraction. A Newtonian prediction of what happens in a particular situation, even a highly stylised one, requires a model. Consider a canon being fired: a model of this includes an idealised representation of the ball and the force of the blast, along with the canon's initial orientation and elevation. The modeller might idealise the shape of the ball as a perfect sphere and assume there is no air resistance. They might make essential use of diagrams like the "free-body diagram". Such a model is *required* for prediction or explanation: theories are too abstract to do that work for any particular circumstance. Other models combine input from *multiple* theories to understand a complex systems like the climate. Climate models draw on thermodynamics, fluid dynamics, and nonlinear dynamics, at least.

Models may also be a means to explore a theory, or to complement one. These uses often occur when the theory is very complicated and difficult to apply in full. Or perhaps the theory leaves open certain questions, which a model fills in for particular cases. Models can also make quantitative what was only qualitative in the theory.

Still other models take no input from theories. Some models are built as part of early theorising, to help scientists develop principles which go on to appear in later theories

(Wimsatt 2007, 104). But some are built to represent target systems in the absence of any theory or desire to develop one. In any case, models always contain more than just the information they (may) inherit from a theory, including diagrams, knowledge about instruments, approximation schemes, and other tools that are not part of any theory (Cartwright 1999).

These various kinds may all exist in ethics too. Here is a stylised background picture which we can use to look for models in ethics. Like scientists, moral philosophers begin with a set of "data": observations of moral life, and our moral judgements. Anti-theorists are right that, like many natural domains, the ethical domain is extremely complex and we have only partial information about it. Ethicists discern certain patterns amongst these data, which they investigate, seeking eventually to systematise them. There may be some empirical regularities (e.g., common judgements, apparent norms), which we aim to explain by the introduction of theoretical concepts (e.g., precisified notions of duty, or welfare). But the domain is complex, patterns are hard to discern, and the data often seem contradictory,[10] and so it is difficult to "read off" moral laws from the data.

The *project* of moral theory nonetheless aims high, seeking moral laws of great scope and generality.[11] This project can be supported by models in ethics as it is in science. One might isolate a particular sub-domain and study it in a model, with the aim of extracting principles which can form part of an ambitious moral theory.

Some criticisms of moral theories highlight the need for models as mediators. Philosophers of science will not be surprised to hear from bioethicists like Arras (2010) that trying to directly apply theories to particular real world cases was unsuccessful. Models can help to connect an existing high-level theory, like utilitarianism, with a particular domain. In science, mediating models bridge this gap by drawing on many elements which are not strictly part of the theory, including empirical information about the domain, approximation techniques, and diagrammatic methods. This accords with how many pieces of ethical work under the framework of a theory look. One's question, or domain of interest, sets the scope of the inquiry. Examples of such domains are distributive questions for social planners, or duties of care. Work done in one of these domains will not usually be expected to apply to the other, even if the philosopher doing the work thinks that a Kantian analysis is best in both cases. This kind of work also involves idealisations, of both the leaving out and distorting kinds. The ethicist studies a situation, or group of people, or situation, which is different in important ways from any real situation. Work of this kind therefore has the indirect nature which is characteristic of modelling. The work might incorporate constraints drawn from real-world considerations and tools ranging from familiar test cases to diagrams and tables. All of these help to connect the content of the theory (a set of principles) to the domain being studied.

But, just as in science, we should also expect to find models operating without theories. One case where this occurs in science is when there is no theory and the scientist works "bottom up" from the phenomena. For ethicists who are skeptical of particular theories, or

[10] All theorists must accept this, I think, as part of their explanation for why so few are adherents to their theory.

[11] In many ways philosophers are *more* ambitious than scientists. In most areas of science there is nothing like the project of moral theory—scientists operate with theories of much more limited scope, and with models. Even in physics, many are sceptical about the prospects for a grand unified theory and spend their time working on domain-specific phenomena. Meanwhile, the discussion in Sect. 2 indicates that ethicists have taken grand unified theory as the natural aim of their discipline.

the project of moral theory as a whole, ethical models might play this role. Another case of relatively theory-free modelling is in mid-level domains which are hard to connect to fundamental theories. Philosophers working on mid-level questions might find it simpler to work directly in the language of their level, rather than seeking connections with the language of the available theories. I take these thoughts up in the next section, where I return to the theory/anti-theory debate.

3.4 Success in Modelling

What counts as success for a model in ethics? It will depend on the purpose of the model. Some models in ethics are machines for rendering judgements about cases. The model is fed a scenario (e.g., described in a vignette about people tied to train track) and it delivers a conclusion which is then tested against "the data"—here, almost always our considered moral judgements.[12] Other models are tested by the quality of the explanations they offer. These focus on rightmaking or goodmaking features. Here, having the right implications is not sufficient; we want the right reasons for those implications. This too is common in science, where there is a large literature on different forms of explanation, and its link to understanding. What is missing in science, and sui generis in ethics, is the link to justification and action. But it is worth noting that in the scientific case the goals of explanation and prediction can come apart, with some models faring well on one and poorly on the other. Perhaps in ethics we shall find models which excel at "getting the answer right" but cannot give us a compelling story about why it is the right answer.

Mathematical models in ethics often translate principles into mathematical constraints. Success here depends on the accuracy of the representation—whether the constraint captures the essence of the principle as it is understood by ethicists who promote it. Often the process of so representing a principle requires the introduction of additional mathematical structure, as I will discuss in Sect. 5 below. In such cases, it is important to perform sensitivity analysis and ensure that the important ethical conclusions do not depend sensitively on assumptions made purely to facilitate the use of the formalism.

Since models have restricted domains of application, it is important to note something which does *not* count as failure: finding a "counterexample" which lies outside the intended domain of the model, or which simply contradicts one of the idealising assumptions. If a model of an inclined plane is developed to study the contribution of gravity, or movement on low-friction surfaces, then it is no rebuttal to point to the existence of a rough ramp, or the differences between the predictions of the model and the behaviour of an object sliding down that ramp. The results from a model are not universal statements, and so not just any counterexample will do. One has to "play the game" and furnish a case which demonstrates that the model fails at its intended purpose, or develop a model which does better—either by having a wider domain, or by generating better results within the common domain.

You can see now why characterising our work as modelling matters. Recognising that moral theorising involves modelling requires a partial re-conception of what it means to

[12] In science, care is taken to separate out which data are used for testing the model. While building a model, the modeller may make use of certain data to calibrate it—ensure it gives the right answers, by adjusting certain parameters. Once the model is ready, it is tested against different data from that used to calibrate it. Success against this new data is taken as confirming the model's usefulness, while success against the data used to calibrate the model is taken to be trivial. I am not sure whether there is a parallel to this in the normative ethics case.

succeed. As the pat phrase goes "all models are wrong, but some are useful". Philosophers are not used to thinking their claims are almost certainly wrong if taken literally. This leads to certain methodological habits which fit poorly with modelling. Most straightforwardly, we cannot take disagreements between models as a sign that one of them must be rejected. Each can be useful for its purpose, so long as those are made clear.

4 Anti-Theory Redux

I now turn to demonstrating some of the benefits of adopting the modelling view, by applying it to three discussions in normative ethics. The first is the anti-theory debate, which opened this essay. The second is drawn from population ethics, and the third from distributive ethics. These are merely examples, chosen based on my interests—my hope is that the reader will be able to apply this view to their favourite bit of ethics by the time they reach the end of the essay.

The theory/anti-theory debate is a dispute about the project of moral theory: whether ethics should be in the business of building theories. In this debate, theories are assumed to have the structure outlined above including, crucially, universality in the scope of laws (or perhaps definitions, in the case of value theory), and with entirely general domains of application. I propose that, insofar as the anti-theory critique does well, it often motivates instead for modelling. I will offer two interpretations of the anti-theory debate in terms of the philosophy of science concepts introduced above. Which interpretation is better will depend in part on what ethicists take themselves to be doing. They need not compete—the first interpretation might better fit some bits of ethical work, and the second, others.

Let us begin by noting that some of the criticisms levelled by anti-theorists at moral theory also apply to scientific theories. Their abstraction and generality makes them hard to work with when explaining specific phenomena, or making predictions about a certain system. Scientific theories, and in particular theoretical laws, present an overly simplified picture of things, such that if one were to make observations "in the wild" one would *not* observe the behaviour predicted by the laws of nature as stated in, for example, Newton's mechanics (Cartwright 1983). But these complaints have no bite against scientific theories—certainly no one would suggest abandoning theorising on the basis of them. One reason for this is that, in science, an important mediating role between theory and world is played by *models* (Morgan and Morrison 1999).

First Interpretation We have ethical theories, and these are precisely the targets of the anti-theory debate: utilitarianism, Kantian deontology, neo-Aristotelian virtue ethics, and so forth. These are abstract and distant from the phenomena that they ultimately describe. They contain laws, which opponents point out often generate the incorrect answer if applied naïvely. Practitioners reply that applying these laws requires skill, which we now interpret as the familiar claim that one must learn how to use models, approximation techniques, and various instruments, to connect these theories with reality.

Second Interpretation We do not (yet) have ethical theories. We are adrift in a complex and confusing domain, and our attempts at systematic investigation should be thought of as modelling in the absence of theory, or modelling which hopes to develop a theory. What is often presented as a law is more like a model-bound regularity whose true domain of application is under investigation.

We begin with the first: there are theories, but we need models to connect them with reality. I think it will be helpful to begin with laws, and how they are thought of by philosophers of science in the modelling tradition I am presenting. To the extent that science involves genuine laws of nature (exceptionless generalities) they are thought to be the laws of physics. But Nancy Cartwright has argued that, even there, laws are best understood as carrying implicit ceteris paribus conditions, and applying "literally" only under abstracted and idealised conditions that are rarely realised in nature (Cartwright 1983). Laws are thus true and serve an important explanatory function, but much of their work is done *through models*. The idealisations in these models serve in part to create situations in which the laws can literally apply. These models don't correspond to exact reality, yet they allow the theory to do its work (Cartwright 1989).

This interpretation offers us a way of understanding and responding to the first anti-theory argument I discussed above: theory simplifies too much. We now see that theories in science are themselves simplified, highly abstract, and distant from the empirical phenomena they purport to explain. Anti-theorists have acknowledged this: Williams acknowledges the parallel but argues that the crucial difference is that in science theories answer to the truth, which allows for scientific theories to be successfully general despite their abstraction. Ethics, for Williams, was inherently local, and he leaned towards non-cognitivism (Fotion 2014, 55–56).

Setting aside the metaethical question, I think Williams neglected the role of models in science and their potential mediating role in ethics. Consider Williams's objections that utilitarianism doesn't reflect the operations that real agents would carry out: the one-thought-too-many objection or the complaint that utilitarianism offers easy verdicts to difficult questions. If we consider a utilitarian model of one of his cases, we should expect that not all of it is intended to correspond to reality. There is nothing methodologically suspect, to the modeller's eye, in claiming that this model is intended to generate successful predictions (i.e., render the correct verdict on the case) but *not* to represent its difficulty. The operations required of the modeller to produce the result may have nothing to do with the cognitive processes by which that verdict would be arrived at by an actual agent.

Baier's argument that seeming laws like "don't kill" are woven into a cultural fabric which provides interpretations, exceptions and specifications now seems like nothing more than Cartwright's analysis of "how the laws of physics lie". Cartwright's claim in the scientific case was that a careful understanding of laws as ceteris paribus generalisations, coupled with close attention to causation, would allow laws to come out as true, and to play an explanatory role in science. (In the next section I will return to the role causation plays for Cartwright and what might be analogous in ethics.)

On this interpretation, the bioethics critique of "high theory" as being unhelpful to that project is correct, unsurprising, and no real challenge to the theories themselves. Whether one works "top down" or "bottom up" in bioethics may depend on how successful one takes moral theories to be *as theories*. If one is independently inclined to think that none of the major moral theories is much good, then one would naturally want to work "bottom up", modelling in the absence of theory—in a way recognisable to any philosopher of biology. If, on other other hand, one thinks that a certain theory is broadly correct, one is more likely to work "top down", using a model to mediate between the theory and real-world cases. All theories need mediating models of this sort.

The second interpretation does better against the other anti-theory arguments I discussed above. On this interpretation we have no theory, and what we call a theory (e.g., utilitarianism) is better understood as a model. This is clearly a more revisionary interpretation, which I don't expect to be attractive to those committed to the project of ethical theory. But it allows those philosophers who are skeptical of that project to find a use for work presented as theory.

Consider the claim that our moral lives contain irremovable moral conflicts or dilemmas, and that "theories" must therefore be false. This is less concerning if we substitute theories for models. Models are false, but hope to be useful. The lack of dilemmas in the model could be a form of idealisation justified as a simplification that is made in order to facilitate analysis. Perhaps the usefulness of the abstraction is then in illuminating the connections between various concepts, or seeing how they work together to generate conclusions. Or perhaps it could be justified as a domain restriction: this is simply a model of cases without moral dilemmas. In those cases, it might be claimed that the model generates the right result.

This no-theory interpretation can also answer Baier's definiteness worry. She claims that the nature of "theories" is such that the norms which feature in them have properties that our actual moral norms do not have. The modeller can here respond that models precisify observed norms into principles for particular purposes, in limited contexts, without claiming that the representation of the actual moral norm in the model is *identical to* or *underlies* that norm. The precision facilitates a certain kind of analysis.

Finally, recall that the very proliferation of contradictory moral theories was taken by Baier as evidence that the project of "theory" cannot, or is at least very unlikely to, succeed. This objection seems tailor-made for the modelling response. It is one of the distinctive features of modelling that we find a proliferation of models which overlap and even contradict one another and yet, in their patchwork fashion, contribute to an overall understanding of their common domain. On this view our different ethical models might be like Teller's two models of water. They fare best in particular areas, explicitly conflict on some questions, and cannot be complete descriptions.

This brings us back to two important features of models discussed above. First, they are purpose-specific, and thus have restricted domains of application. Second, this means that they are not sensitive to counterexamples in the way that fully general theories are.

What could it mean to say that utilitarianism, say, has a particular purpose or restricted domain? These domains could be types of question, as I will discuss below for prioritarianism, or something as general as Nozick's "push" and "pull" factors for morality (Nozick 1981). However they are spelled out, the result will be that certain questions simply aren't meant to be addressed by the model. This may (and probably will!) seem unsatisfactory to the ethicist used to debate by counterexample. If a theory is doing poorly in the general case, why trust it in a limited domain? We are rightly suspicious of a theory that says you can sometimes torture children and should feel uncomfortable about using it in non-child-torturing situations![13]

There are three parts to the modeller's reply. The first is simply to insist that we are working in a complex, contradictory domain. Remember that on this interpretation, *we do not have theories* in the sense introduced above. All of the available models are limited, and face "counterexamples", be they child torturing or Nazis at the door. The second part of the response is to articulate, in a non ad hoc way, a domain restriction. This will be determined by the purpose of the inquiry in which the model features, and the idealisations built into the model. Contrary data only *fails* to be a counterexample if it is genuinely outside of the model's domain. "It is just a model" cannot be a Get Out of Jail Free card;

[13] Thanks to Gustaf Arrhenius for pushing me on this point.

it ought to be a description of a careful and principled methodological approach. Third, the modeller notes that we can still have conflicts between models and judge one better than the other. Consider one model, with a particular purpose and associated domain, outside of which it advocates for torturing children. Now consider a second model which has a wider domain—it can answer the same questions as the first model, and more. On the common domain, the second model does as well as the first. The second model's wider domain includes the child-torturing cases, and it does not deliver the same incorrect result. In that case, the first model is clearly worse than the second. Worse for what? Well, for all purposes the two models have in common, and for general use—given that they perform equally well on common questions, having wider scope is desirable as it unifies and simplifies inquiry.

4.1 Moral Particularism

The foregoing discussion also gives us a way to think about moral particularism. Particularism is something like the view that ethical theory impossible *and* that there is no middle-ground whatsoever. We must confront particular cases in all their granularity, rather than attempt any systematisation (e.g., Dancy 2017). Modelling seems to offer us a way to access precisely the middle-ground that particularists deny, however. It makes no claim to universality, or general application. Models can be local, they can synthesise only some of the available data. Importantly, they need not be axiomatisable, or decidable, or even formal. They are the tools of scientists engaged in the sort of ground-up work particularists seem to want us to engage in, but they achieve more in the way of generality than they take to be possible.

Writing about normative models in decision theory, Michael Titelbaum comments thus on particularism: "The normative modeler proceeds piecemeal, trying to solve local problems and gradually extend the boundaries of normative knowledge. (In this she is much like the working scientist.) The modeler does not fully yield to the particularist's insistence on treating each case on its own terms, but neither does she assume that the normative is a single, systematizable domain" (Titelbaum forthcoming, 16).

The hard-line particularist will reply that this is doomed to fail because it assumes that moral considerations function the same way across circumstances. For example, consider Dancy's reasons holism, under which a feature of a situation, like the fact that someone is lying, can have a different moral *valence* across cases (Dancy 2017, S3). We can situate this on our philosophy of science map by returning to Cartwright on laws of nature. In her more controversial work, Cartwright claims that laws of nature are literally false in much the way that particularists claims that the maxims in ethical theories are false. So, what justifies the use of laws, according to Cartwright? She argues that using laws requires the postulation of capacities, which *act in the same way in all circumstances*, despite the apparent falseness of the lawlike statements of science.

> The logic that uses what happens in ideal circumstances to explain what happens in real ones is the logic of tendencies or capacities. What is an ideal situation for studying a particular factor? It is a situation in which all other "disturbing" factors are missing. And what is special about that? …This tells you something about what will happen in very different, mixed circumstances—*but only if you assume that the factor has a fixed capacity that it carries with it from situation to situation.* (Cartwright 1989, 190f, my emphasis) quoted in (Reutlinger et al. 2019)

In science these capacities are causal powers, which clearly won't do for ethics. Under our analogy, the reasons holist denies that morality has anything analogous to nature's capacities. The aspirant theorist thinks that it does, that there are goodmakers and rightmakers which have the same action across situations. This clarifies what the debate is about. The mere fact that moral laws don't straightforwardly apply in observed cases is not an argument in favour of particularism, as it does not undermine the existence of these capacities. The case for particularism must be a case against constant capacities—constant goodmakers and rightmakers.

Models, recall, have limited scopes set by the purposes of the modellers and the idealisations they employ. One factor which might set the scope of an ethical model is the modeller's beliefs about the constancy of action of goodmakers and rightmakers. There is a spectrum of possible models. At one end, we might model a set of paradigm cases, in order to generate a limited-scope principle which applies only locally. At the other, our models will be stepping stones to a fully general theory with universal principles.

5 Impossibility Theorems in Population Ethics

In this section I want to observe some idealisations in population ethics, and comment on how my modelling view might illuminate results in that field.

Population ethics is the study of ethical problems concerning populations—groups lives, people living for a given time with a given level of welfare. It is considers actions which affect how many people will live at a future time, and which people they will be. Amongst other things, it seeks a population axiology; that is, an ordering of populations with regards to their (intrinsic) goodness (Arrhenius forthcoming). It often proceeds by thinking about which of two possible populations is better. The standpoint in population axiology is not one of considering action, for example bringing each population into being, but rather a judgement of their relative goodness. Following Derek Parfit's presentation of his "mere addition paradox", it has been recognised that there are significant difficulties in formulating such an ordering (Parfit 1984, Ch.19).

One popular strand of population ethics focuses entirely on welfare. (Conceived, very roughly, as how well a person's life is going; how good it is for them.) It is in this context that various paradoxes and associated impossibility results arise. One might think that this is a problem only for welfarists, but Gustaf Arrhenius pushes back against this restriction:

> Since *we can assume that other values and considerations are not decisive* for the choice between the populations above, as we shall show below, this is not true. Hence, paradoxes like the above are a problem for all moral theories which hold that *welfare at least matters when all other things are equal*. Since, arguably, any reasonable moral theory has to take this aspect into account when determining the normative status of actions, the study of population ethics is of general import for moral theory. (Arrhenius forthcoming, 5, emphasis mine)

As Arrhenius puts it, the focus on welfare is not because other considerations—such as fairness, liberty, and virtuousness—do not matter. They may well figure in the ranking of populations. But the population ethicist assumes "that welfare at least matters when all other things are equal". This is a clear idealisation—an omission of these other factors, on the grounds that they are being assumed to be equally balanced in the weighing of considerations. Put another way, it is a ceteris paribus clause. As we've just seen, these play a crucial role in the strand of philosophy of science I am drawing on.

How are we to interpret the results of population ethics, given this idealisation? One of Arrhenius's contributions is to present precise theorems showing the impossibility of satisfying various conditions which are taken to be necessary features of an adequate population axiology. He proceeds by first introducing such a condition informally, on the basis of intuitive responses to cases. For example, avoiding the Repugnant Conclusion is one condition of adequacy. This is the result that, for a possible population of many high-quality lives, there is some much larger population of people living lives barely worth living, which is ranked *better* than the former by the population axiology. In general, Arrhenius's method is to first formulate an adequacy condition in words, on the basis of the relevant intuition-eliciting case or reflection, and then to introduce an exact formulation which employs mathematical representations.

Here is an example of a condition which is part of the precisification of avoiding the Repugnant Conclusion.

> *Quality*: There is a perfectly equal population with very high positive welfare which is at least as good as any population with very low positive welfare, other things being equal.
> *Quality (exact formulation)*: There are two positive welfare ranges $R(u, v)$ and $R(1, y)$, $u > y$, and a population size $n > 0$, such that if $W_z \subset R(u, v)$, $A \subset W_z$, $N(A) = n$, and $B \subset R(1, y)$, then A is at least as good as B, other things being equal. (Arrhenius forthcoming, 304).

This seems to me to clearly be a model.In addition to the basic feature that Arrhenius is employing mathematical representations to make his arguments precise, I note two other characteristic features of models: (1) "structural" assumptions are introduced to facilitate the mathematical representation, and (2) idealisations are introduced to simplify the analysis.

We have already discussed one idealisation involved: the focus on welfare. As an example of a structural choice, Arrhenius uses sets to represent welfare levels and he assumes that the set of welfare levels is fine-grained, in the following sense (Arrhenius forthcoming, 299):

> *Finite Fine-grainedness*: There exists a finite sequence of slight welfare differences between any two welfare levels.

So, what are we to make of the fact that this work involves modelling? Arrhenius presents his work as illuminating something about the structure of value, or of our intuitions about value. He is careful in his conclusions:

> If the evaluations above stand up to scrutiny, that is, if we find it impossible to give up any one of them, then our considered moral beliefs are mutually inconsistent. And if consistency with considered intuitions is a necessary condition for a moral theory to be justified, we seem to be forced to conclude that there is no such theory which can be justified. In other words, paradoxes of the above kind might challenge some of our deepest beliefs about moral justification and the meaningfulness of moral theories. (Arrhenius forthcoming, 4)

So, if these are the data, and fitting all the data is a requirement for a theory, then there is no moral theory (Arrhenius 2000, forthcoming). Here I would make but a friendly amendment: If these are the data, and fitting all the data is a requirement, *and this model—with* *its idealisations and structural assumptions—tells us something general about value*, then there is no moral theory. The italicised addition is crucial.

As Cartwright shows, when other things are not equal, modelling is much more difficult than in the ideal case. In Cartwright's picture, the movement out of the idealised case is licensed by nature's capacities acting in fixed ways from situation to situation. Modellers must engage in careful work to get their results to apply in messy real situations, either by sensitivity analysis, de-idealising the model, or presenting their results with explicit provisos linking them to the assumptions under which they were generated.

Now, let us suppose that Arrhenius's results *do* show that we can have no consistent *theory* of value, which captures all of this data. The population ethicist need not despair. There are many domains of science in which we have no overarching theory, or where we know that two successful models of sub-domains cannot be unified in a consistent manner. Fundamental physics is just such a case, where quantum field theory and general relativity, each highly successful in its domain, cannot currently be made consistent.

The modelling strategy is to go local, and construct models which capture some of the data, in some circumstances. As modelling is purpose-driven, this may require population ethics to become more applied. By responding to real-world problems, population ethicists may be able to reject an assumption, or to prioritise which of the conditions of adequacy are most important. This sort of purpose-driven prioritisation would then motivate the construction of a more local model of a population axiology—one which is known to be incomplete, but which can still be useful.[14]

6 Models of Prioritarianism

In distributive theory, philosophers discuss the plausibility of distributive principles with respect to short vignettes presenting cases. This is another clear case in which I see modelling at work.[15] Here we face the same choice as in Sect. 4, of regarding distributive theories like prioritarianism as mere models, or of characterising them as theories which make contact with particular cases through models of the theory.

In much distributive theorising, the distribution problem is summarised in a table, containing a numerical representation of the distribution problem. Here is a classic case in which Derek Parfit presents a case due to Thomas Nagel.

> Nagel imagines that he has two children, one healthy and happy, the other suffering from a painful handicap. He could either move to a city where the second child could receive special treatment, or move to a suburb where the first child would flourish. […then, quoting Nagel:] I want to suppose that the case has the following feature: the gain to the first child of moving to the suburb is substantially greater than the gain to the second child of moving to the city. […] To ask my questions, we need only two assumptions. First, some people can be worse off than others, in ways that are morally relevant. Second, these differences can be matters of degree. To describe my imagined cases, I shall use figures. Nagel's choice, for example, can be shown as follows. (Parfit 2002, 81–83)

[14] This is similar to the approach taken by Budolfson and Spears (2022), although their approach is to reject one adequacy condition outright rather than to neglect it for heuristic reasons.

[15] Thanks to Nic Côté for suggesting this as a case study.

Table 1 Two-child case, from Parfit (2002, 83)

	First child	Second child
City	20	10
Suburb	25	9

Table 1 reproduces his table.

There follows this passage, explaining the table.

> Such figures misleadingly suggest precision. Even in principle, I believe, there could not be precise differences between how well off different people are. I intend these figures to show only that the choice between these outcomes makes much more difference to Nagel's first child, but that, in both outcomes, the second child would be much worse off. One point about my figures is important. Each extra unit is a roughly equal benefit, however well off the person is who receives it. If someone rises from 99 to 100, this person benefits as much as someone else who rises from 9 to 10. Without this assumption we cannot make sense of some of our questions. We cannot ask, for example, whether some benefit would matter more if it came to someone who was worse off. (Parfit 2002, 83)

This example is very naturally interpreted as modelling. The case presented in the vignette is prepared for analysis by representing it formally. The story contains no numbers, they are introduced as a thinking aid, along with some particular interpretative principles. The immediate object of analysis is what we might call the benefit structure, displayed in the table. Importantly, this structure has features which we are told to to disregard—in particular, precise comparability. This precision is an artefact introduced by a choice made by the modeller, Parfit, because they want to use other features of the numerical structure. This way of representing cases becomes a framework for modelling principles of distributive justice. The views under discussion, equality and priority, are rendered as principles about the numbers in the table. If we interpret egalitarianism as a theory, then what Parfit creates is a model of the theory within this framework. For example, egalitarianism, a view about people being equally well off, becomes a view about equality between numbers representing people's welfare. Finally, we see the signature feature of modelling in the subsequent discussion in the literature: these benefit structures and the models of principles are investigated and discussed as proxies for the substantive views about justice.

Interpreting this as modelling is not merely possible, it is helpful. To show how it might help us make progress, I want to consider a more recent debate about prioritarianism. It begins with Otsuka and Voorhoeve (2018), who argue against a form of prioritarianism and in favour of a form of egalitarianism. They use a similar case to the above, introducing some additional structure which is worth reflecting on. In their case, there is uncertainty about the outcome (in the form of objective, given probabilities). Otsuka and Voorhoeve then make the following qualifications.

> We shall assume a measure of utility on which a prospect has higher expected utility for a person just in case it would be preferred for that person's sake after rational and calm deliberation with all pertinent information while attending to her self-interest only. (A person's expected utility is just the probability-weighted sum of her utility in each possible state of the world.) One prospect has the same expected utility as another for a person just in case such deliberation would yield indifference between the two prospects.

> [In a footnote to the above:] In other words, we assume that the measure of utility is derived from idealized preferences satisfying the Von Neumann-Morgenstern axioms.[…] More generally, throughout, we assume that orthodox decision theory applies, according to which under risk, a decision-maker ought to maximize the expectation of what he takes to be the relevant value (so that a utilitarian ought to maximize the sum-total of expected utility, a final-utility prioritarian the sum-total of expected priority-weighted utility, etc.). (Otsuka and Voorhoeve 2018, 9, fn.7)

Here the model is fitted with additional structure, to facilitate yet more precise analysis. The distributive theories being discussed (prioritarianism and egalitarianism) do not involve in any essential way these views on utilities, their measurement, and their relation to decision theory. The decision-theoretic link is particularly interesting: decision theory is itself a model (Roussos forthcoming); in particular a representational model of agents, which employs various distorting idealisations. Some of these, like the transitivity of preference, are normative assumptions. So, if VNM agents differ from real agents like you and me in this regard, the explanation of that difference is that we ought to be like them. But some of the idealisations are not normative: e.g., these agents have complete preferences. I understand this as a heuristic idealisation: decision theorists know it is not true, but it is included to simplify the analysis by facilitating the use of certain mathematical structures.

Otsuka and Voorhoeve's model also goes beyond the VNM model, by making comparisons between the utilities of individuals possible. Its conclusions are therefore a complex result of non-normative idealisations about agents, normative idealisations concerning those agents' rationality, additional assumptions to achieve interpersonal comparison of these utilities, and assumptions about how the principles under discussion (prioritarianism and egalitarianism) are realised in the model.

What might it mean to say that a model of this sort has a restricted domain of application? To illustrate this, consider the argument that Otsuka and Voorhoeve make against prioritarianism. They consider a variant of the case with just one child, and uncertainty about whether the child will become disabled, and thus how it might benefit from a move to the suburbs or city. Otsuka and Voorhoeve argue that prioritarianism treats risky intrapersonal trade-offs like this as "involving the very same moral calculus as interpersonal trade-offs in which the interests of different people conflict" (Otsuka and Voorhoeve 2018, 9). In so doing, it fails to appreciate the unity of individuals.

A prioritarian might respond that such a case is simply outside of the scope of this model. Prioritarianism, as a model of the good, embodies what an ideally virtuous agent would desire in contexts of *impartiality*, reflecting its primary intended application to questions of distributive social policy. Its domain is multi-person cases, because it is intended to answer the question of how we weigh the competing interests of different individuals.

This is similar to a defence of prioritarianism offered by (Adler and Hotug 2019). But rather than argue for this as a restriction of the domain of their model, they claim that this is simply the domain of *ethics itself*. "Morality is a framework for resolving interpersonal conflicts; but in a one-person universe there can be no such conflicts" (Adler and Hotug 2019, 121). They do offer a justification of this restriction in scope which seems well-suited to my modeller's answer above: "prioritarians can invoke Otsuka and Voorhoeve's favored explanation of the difference between interpersonal and intrapersonal conflicts, namely an appeal to respectively the separateness and the unity of persons, to motivate such a restriction in scope. According to this line of argument, the unity of persons is decisive in one-person cases…In cases of interpersonal conflict, on the other hand, the separateness of persons comes into play and motivates a prioritarian weighting" (Adler and Hotug 2019, 121). I.e., the prioritarian model has a natural domain of interpersonal conflict.

7 Conclusion

This paper advertises a certain way of seeing ethical work: as modelling akin to that in science. I have shown what it means to think of ethics as modelling: it involves idealisations deployed in representations which are used for indirect or proxy inquiry. This interpretation has a number of benefits. Chief among them is the commonality with scientific methodology—not because there is anything special about science but because there happens to be a decades-long methodological literature on how modelling succeeds in that domain. Ethicists can draw on this to manage their own use of idealisations, and improve their inferences from model to target. In this concluding section I offer a few final methodological suggestions for those who see themselves in my descriptions.

Before I come to those suggestions, let me make two comments about the scope of this essay. First, I do not claim that all of ethics is modelling. I am guided by the characteristics I introduced in Sect. 3: modelling involves indirect inquiry using idealised representations. I take seriously those ethicists who say that they are offering theories, and that their laws are of perfectly general scope. Other non-modelling work includes work which focuses on defining terms, or which makes conceptual distinctions, or which addresses cases but without the use of representations and proxy systems.

Second, I acknowledge that I have offered little in the way of a defence of the way of working that I have here characterised as modelling. I have noted but a few benefits of modelling, such as its tools for managing idealisation and its ability to facilitate checking a set of principles for coherence. No attempt has been made to compare it to non-modelling methods, or to defend it over those methods. So, for skeptics of this way of working, my interpretation of it as modelling won't furnish any additional reasons to adopt that method. (Though I hope it may offer understanding of what is happening in this way of working and why.) In future work, I hope to develop such a sustained defence of modelling in ethics.

Those caveats aside, let us return to methodology. What should one do differently, as a modeller? First, ethicists who are modellers will want to place more emphasis on articulating the purposes of their inquiry, linking these to the idealisations and structural assumptions they employ, and noting the consequent restrictions and caveats attached to their results. This is crucial because modelling is so purpose-specific, and the evaluation of models is on their usefulness for their intended purposes. As I noted above, some of this happens in ethics today, but in a piecemeal and unsystematic way.

Second, ethicists who adopt modelling will need to re-evaluate the role of counterexamples. As a practice, moral philosophy currently thrives on the generation of principles, and their testing against and adjustment in the face of counterexamples, typically in the form of stylised cases where our intuitions contradict the recommendation of the principle. Because models are domain specific, and justified by particular purposes, many seeming "counterexamples" have no bite against them. Thus, new ways of arguing will need to be developed, and old ones revised or abandoned.

Third, practices for working with idealisations should be developed or refined. I briefly discussed sensitivity analysis, the careful variation of assumptions to observe the dependency of the result on small changes. This is a crucial method in scientific modelling that I do not observe widely used in ethics. It is crucial for producing results which do not depend sensitively on idealising assumptions that the modeller knows to be false.

All of this may involve new foci in education: rather than being trained to seek counterexamples, graduate students in ethics might be trained in the careful use of idealisation, the analysis of its effects through sensitivity analysis, and the critique of opposing models. This will bring with it a new comfort with making progress locally, either toward an eventual goal of theory or in the absence of any such endpoint.

A final thought. In my discussion of the anti-theory critique, I have offered a new way of understanding fundamental ethical "theories", such as utilitarianism, Kantianism, and virtue ethics. This is to understand them as models. Such an interpretation would significantly curtail their ambitions, and would involve specifying purposes, idealisations, and domains of applicability for each of them. The value of adopting such an approach is that it furnishes us with a new set of tools for understanding the disputes between these models, and for working with multiple models.

Acknowledgements Thanks to Krister Bykvist for in-depth discussions and feedback on an earlier version of this paper, to Nicolas Côté for being an enthusiastic audience and friendly critic to the earliest versions of these ideas, and to Alex Voorhoeve and Orri Stefánsson for their constructive feedback. The reviewers of this journal were of particular help in clarifying the nature of my interpretative claim; my thanks to them. For helpful comments and criticisms, thanks also to participants of the IFFS PPE seminar, Stockholm Practical Philosophy Higher Seminar, and LSE seminar on distributive justice and social choice. This work was conducted in the research project Climate ethics and future generations, funded by Riksbankens Jubileumsfond (grant number M17-0372:1).

Declarations

Conflict of Interest The author declares that they have no conflict of interest.

References

Adler MD, Hotug N (2019) Prioritarianism: a response to critics. Politics Philos Econ 18(2):101–144

Anscombe GEM (1958) Modern moral philosophy. Philosophy 33(124):1–19. https://doi.org/10.1017/S0031819100037943

Arras J (2010) Theory and bioethics. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy. Winter 2016. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/theory-bioethics/. Accessed 08 May 2021

Arrhenius G (2000) An impossibility theorem for welfarist axiologies. Econ Philos 16(2):247–266. https://doi.org/10.1017/S0266267100000249

Arrhenius G (forthcoming) Population Ethics. Oxford University Press

Baier A (1985) Postures of the Mind. University of Minnesota Press, Minneapolis

Baier A (1989) Doing Without Moral Theory. In: Clarke SG, Simpson E (eds) Anti-Theory in Ethics and Moral Conservatism. State University of New York, Albany, pp 29–49

Beck L, Jahn M (2021) Normative models and their success. Philos Soc Sci 51(2):123–150

Bovens L, Hartmann S (2003) Bayesian Epistemology. Oxford University Press, Oxford

Budolfson M, Spears D (2022) “Does the Repugnant Conclusion Have Important Implications for Axiology or for Public Policy?” In: Oxford Handbook of Population Ethics, edited by Tim Campbell, Krister Bykvist, and Gustaf Arrhenius. Oxford University Press, Oxford

Cartwright N (1983) How the Laws of Physics Lie. Oxford University Press

Cartwright N (1989) Nature’s Capacities and Their Measurement. Oxford University Press

Cartwright N (1999) The Dappled World: A Study of the Boundaries of Science. Cambridge University Press, Cambridge

Chappell S-G (2022) “If Not Moral Theory, Then What?” In *Epiphanies*, 49–110. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780192858016.001.0001

Clarke SG (1987) Anti-theory in ethics. Am Philos Q 24(3):237–244

Dancy J (2017) Moral particularism. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy. Winter 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/moral-particularism/. Accessed 26 Sept 2020

Eva B, Hartmann S (2020) On the origins of old evidence. Australas J Philos 98(3):481–494. https://doi.org/10.1080/00048402.2019.1658210

Fotion N (2014) Theory vs. Anti-Theory in Ethics: A Misconceived Conflict. Oxford University Press

Frigg R, Hartmann S (2018) “Models in Science.” In: The Stanford Encyclopedia of Philosophy, edited by Edward N. Zalta, Summer 2018. Metaphysics Research Lab, Stanford University

Glymour CN (1999) “Realism and the Nature of Theories.” In: Introduction to the Philosophy of Science, edited by Merrilee H. Salmon. Indianapolis and Cambridge: Hackett

Godfrey-Smith P (2006) Theories and models in metaphysics. The Harvard Review of Philosophy 14(1):4–19

Godfrey-Smith P (2012) Metaphysics and the philosophical imagination. Philos Stud 160(1):97–113

Hancox-Li L (2017) Idealization and abstraction in models of injustice. Hypatia 32(2):329–346

Hempel CG (1966) Philosophy of Natural Science. Foundations of Philosophy. Prentice Hall, London

Leitgeb H (2013) Scientific philosophy, mathematical philosophy, and all that. Metaphilosophy 44(3):267–275

List C, Valentini L (2016) The methodology of political theory. In: Oxford Handbook of Philosophical Methodology. Oxford University Press, Oxford. http://personal.lse.ac.uk/list/PDF-files/MethodologyPoliticalTheory.pdf. Accessed 25 Sept 2020

Louden RB (1990) Virtue ethics and anti-theory. Philosophia 20(1):93–114. https://doi.org/10.1007/BF02382586

Louden RB (1992) Morality and Moral Theory: A Reappraisal and Reaffirmation. Oxford University Press, New York

McCarthy D, Mikkola K, Thomas T (2019) Aggregation for potentially infinite populations without continuity or completeness. http://arxiv.org/abs/1911.00872. Accessed 24 Sept 2020

McKeever S, Ridge M (2015) Obvious Objections. Oxford University Press

Morgan MS, Morrison M (eds) (1999) Models as Mediators. Cambridge University Press, Cambridge

Novick A (2017) Metaphysics and the vera causa ideal: the nun’s priest’s tale. Erkenntnis 82(5):1161–1176. https://doi.org/10.1007/s10670-016-9863-1

Nozick R (1981) Philosophical Explanations. Clarendon Press, Oxford

Nussbaum M (2000) “Why Practice Needs Ethical Theory.” In: Moral Particularism, edited by Brad Hooker and Margaret Olivia Little, 234–345. Clarendon Press, Oxford

O’Neill O (1987) “Abstraction, Idealization and Ideology in Ethics.” In: Moral Philosophy and Contemporary Problems, edited by J. D. G. Evans, 55–69. Royal Institute of Philosophy Supplements 22. Cambridge University Press, Cambridge

Otsuka M, Voorhoeve A (2018) Equality versus priority. In: Oxford Handbook of Distributive Justice. Oxford University Press, Oxford. http://personal.lse.ac.uk/OTSUKAM/M. Accessed 25 Sept 2020

Parfit D (1984) Reasons and Persons. Clarendon Press, Oxford

Parfit D (2002) Equality or Priority? In: Clayton M, Williams A (eds) The Ideal of Equality. Palgrave Macmillan, New York, pp 81–125

Paul LA (2012) Metaphysics as modeling: the Handmaiden’s tale. Philos Stud 160(1):1–29

Potochnik A (2017) Idealization and the Aims of Science. University of Chicago Press, Chicago

Raphael DD (1974) The standard of morals: the presidential address. Proc Aristot Soc 75:1–12

Reutlinger A, Schurz G, Hüttemann A, Jaag S (2019) Ceteris paribus laws. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy. Winter 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/ceteris-paribus/. Accessed 26 Sept 2020

Roussos J (forthcoming) Normative formal epistemology as modelling. Br J Philos Sci. https://doi.org/10.1086/718493

Suppes P (1969) A Comparison of the Meaning and Uses of Models in Mathematics and the Empirical Science. In: Suppes P (ed) Studies in the Methodology and Foundations of Science. Reidel, Dordrecht, pp 10–23

Teller P (2001) Twilight of the perfect model model. Erkenntnis 55(3):393–415

Timmons M (2012) Moral Theory: An Introduction. Rowman & Littlefield Publishers, Lanham

Titelbaum MG (forthcoming) "Normative Modelling." In: Methods in Analytic Philosophy: A Contemporary Reader, edited by J. Horvath. The PhilPapers Foundation

Weisberg M (2007a) Three kinds of idealization. J Philos 104(12):639–659

Weisberg M (2007b) Who Is a modeler? Br J Philos Sci 58(2):207–233

Williams B (1981) Moral Luck: Philosophical Papers 1973–1980. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139165860

Williams B (1973) "Critique of Utilitarianism." In: Utilitarianism: For and Against, edited by Bernard Williams and J. J. C. Smart. Cambridge University Press, Cambridge

Williamson T (2018) Model-Building as a philosophical method. Phenomenol Mind 15(15):16–22. https://doi.org/10.13128/Phe_Mi-24968

Williamson T (2006) "Must Do Better." In: Truth and Realism, edited by Patrick Greenough and Michael P. Lynch, 177–88. Clarendon Press; Oxford University Press, Oxford: New York

Williamson T (2017) "Model-Building in Philosophy." In: Philosophy's Future: The Problem of Philosophical Progress, edited by Russell Blackford and Damien Broderick. Wiley, Oxford

Wimsatt WC (2007) Re-Engineering Philosophy for Limited Beings. Harvard University Press, Cambridge