Modelling in moral philosophy

Joe Roussos

October 2020

Note: This is a work in progress, and a new venture into moral philosophy for me. I previously wrote about how decision theory and formal epistemology are forms of normative modelling, focusing on their formal (i.e., mathematical) natures. This essay represents an expansion of the idea that philosophers are engaged in modelling into a much wider domain of normative and evaluative thinking and with fewer attachments to mathematical presentation. But, I am not a moral philosopher! So if this project is to succeed I need help from ethicists and metaethicists. Comments, thoughts,and challenges are thus even more welcome and encouraged than is usually the case.

1 Introduction

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

What is a model? I shall spend some time on this, but here is a brief example to get us started. In order to study the population of fish in my local pond, I observe the fish feeding, breeding, and dying, for a few generations. I realise that the pond has a finite capacity for fish, due to their needs for space and competition for food. I observe that the population this week generally depends positively on the population last week, but that as the population reaches the capacity of the pond, crowding hampers population growth. In order to predict the population behaviour, I decide to use the following equation: $N_{t+1} = 4N_t(1 - N_t)$, where N is the number of fish in the pond divided by the carrying capacity, and t is a time index counting weeks.

In so doing, I am modelling the fish population. There are a few notable features of this model. I have represented some aspects of the fish population mathematically. In so doing, I have ignored many features of the real pond and fish, such the natural variation in fish size and reproduction. I am treating time as discrete, and counting only weeks. I have ignored certain factors which I know to influence the population level, such as fishing. I make no claims that this equation describes fish growth everywhere: the 4 is a parameter that I choose based on my local observations. These are characteristic features of modelling as a method.

Talk about "modelling" as a method of philosophical inquiry is increasingly prevalent, across various philosophy subfields. Williamson (2006, 2017) has named modelling as important method for a certain style of philosophy; what we might call scientific or mathematical philosophy. In formal philosophy of science, epistemology, and decision theory, it is increasingly common for philosophers to describe their own practice as modelling (Bovens and Hartmann, 2003; Eva and Hartmann, 2019), and to advocate for it as a method (Leitgeb, 2013, p. 273).

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 advocating for the adoption of new scientific methods.

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, does, and should make use of idealised models like my model of the fish population in my pond.

Several things come out of this methodological claim. The first is a new way of understanding what is going on (and going wrong) in the theory/anti-theory debate in ethics. It also gives us a new way of understanding impossibility theorems in population ethics, and their bearing on ethics as a whole. Finally, the fact that ethicists have (unknowingly) been modelling comes with certain methodological constraints for them. Most notably, models are not sensitive to counterexamples in the way that much of ethical theory is taken to be. If I am right, this requires a significant shift in how ethicists practice their craft.

I begin in section 2 with some motivation, considering the anti-theory debate as an entry-point for my analysis. Section 3 sketches my idea, that normative ethics involves modelling. We then turn to science: section 4 gives a more complete description of what a model is, and section 5 considers modelling as a methodology. With this in place, we can return to normative ethics in section 6, where I characterise a moral theorising as modelling in more detail, reflecting on the anti-theory debate, distributive theory in particular, and population ethics. Section 7 concludes with a manifesto for modelling in normative ethics.

2 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" critique. 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.[1] A moral theory:

- provides a decision procedure for determining which actions are right or wrong (Chappell, unpublished; Fotion, 2014, p. 40; Louden, 1992, p. 97; Timmons, 2012, p. 13; Williams, 1981, pp. ix–x).
- is axiomatisable; i.e., can be stated by theorists in terms of a finite set of principles (Fotion, 2014, p. 41; Louden, 1992, p. 97).
- is decidable; i.e., one can check whether any particular action or belief is correct according to the theory (Fotion, 2014, p. 40; Nussbaum, 2000, p. 234).
- 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, unpublished; Fotion, 2014, p. 44).
- is or aspires to be the uniquely true theory. Theories compete; there is only one correct theory (Chappell, unpublished; Fotion, 2014, p. 42).[2]

Some combination of the above is often taken to mean that there are no moral dilemmas (Louden, 1992, p. 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, p. 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] But there is, of course, a spectrum of systematicity and ambition. Rawls's *Theory of Justice*, which is the paradigm example of "ideal theory" in political philosophy, restricts itself to justice and indeed to articulating only some aspects of justice-as-fairness, leaving aside other normative questions.

[1] Many disagree that "theory" has these properties or ought to have them. That is fine, I aim only to characterise an existing debate, in order to find an entry point for my methodological analysis. As will become clear, I am in no sense endorsing this characterisation.

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

[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.

[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).

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.

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, p. 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.

Linked to this is the claim that our moral lives *do* contain irremovable moral conflicts or dilemmas. If theories demand that "there be one and only one morally correct answer, justified by norms" then theory which captures these dilemmas is simply impossible (Clarke, 1987, p. 239). Thus theory cannot but fail at capturing the full reality of the moral domain.

Second, the very proliferation of moral theories shows that they are unlikely to succeed. Annette Baier claims that 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, pp. 33–34).

Third, 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 in determining their application" (Clarke, 1987, p. 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 may be something right to the anti-theory critique of certain overly-ambitious bits of ethical theory. But there is something deeply wrong in how some anti-theorists characterise the space of methodological options. They seem to take the options to be "theory" (bad), or a form of very granular 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. One goal of this paper is to explore that middle-ground and advocate for its use.

I will begin with a brief description of the idea that normative ethics can, does, and should use models. I will then describe in more detail what modelling is in science, before coming back to moral philosophy. With target methodology in place, I will examine different examples of normative ethics that might "fit the bill".[5]

3 The suggestive idea

The ethical domain is extremely complex and we have only partial information about it. This complexity means that we cannot read off moral laws from the data (be they observed moral behaviours or our considered moral judgements). Nor can we formulate a full and precise description of them which is also tractable for analysis. Instead we must simplify, idealise, and examine only one part of the ethical domain at the time. Our efforts are partial, and distorted by the accommodations we make to our own limitations. In short, we build models.

Here are three versions of this claim, in order of increasing ambition.

- Least: there exist some instances of modelling within ethics. The easiest cases are formal work, incorporating explicit idealisation, performed by philosophers with mathematical and scientific training that leads them to speak in terms of models and modelling. For example, (McCarthy, Mikkola, and Thomas, 2019) and (Colyvan, Cox, and Steele, 2010).

- Moderate: modelling is widespread in formal and semi-formal ethics, including work on value theory, welfare economics, distributive theory, population ethics, and moral uncertainty. This includes work done by philosophers who would say they are *not* modelling and who intend their work to have universal scope.

- Most: our conflicting intuitions and failures to achieve agreement on consistent systematising moral theories are indications that this is a complex domain with limited information. Therefore moral theories, which are typically framed as having universal scope, ought to be thought of instead as models. The framing of ethical principles, whether criteria of rightness or goodness, in universal language represents a methodological mistake, overreaching what can be achieved by creatures with our limited means and limited understanding. We are all modelling, or we ought to be.

I believe the most ambitious version of this claim. I also think there is nothing wrong with this. As we attempt to theorise the ethical domain, we make progress

[5]Note again the focus on normative ethics. Our subject here is not representational models of communities obeying norms, or of how norms might emerge, such as those studied in the literature on the social evolution of morality. I am here interested in models whose purpose is the generation of normative claims.

in much the same manner as scientists investigating their complex domains, be they as broad as population ecology and the climate, or as seemingly simple as a note of currency floating in the wind.

It does require a re-conception of what it means to succeed at moral theorising, however. As the pat phrase goes "all models are wrong, but some are useful". Philosophers are not used to thinking their claims are certainly wrong, which leads to certain methodological habits which we need to unlearn.

Let me say at the outset that conceiving of normative ethics as a modelling discipline need not undermine the traditional project of establishing universal moral laws. As we will see sciences which aim to generate laws, such as physics, make repeated use of models as mediating tools to bridge between theory and reality.

With my target on the table, let us look in more detail at what models are and how they work in the sciences.

4 Scientific models

Let us begin with a review of scientific models, and the methodological lessons we have learned from five or six decades of philosophical study of modelling.[6]

"Model" is one of those unhelpful terms that is used to mean many different things, so I want to begin with a common meaning that I *do not* intend to use: the meaning logicians give to the term. Roughly put, logicians use "model" to mean an interpretation that satisfies a set of sentences. An interpretation is here an assignment of semantic values to the basic vocabulary in use. This semantic sense of "model" takes it to pick out certain *mathematical structures*. Some philosophers of science (e.g., Suppes (1969)) have argued that this meaning of "model" is the same as, or should be used to explicate, the workaday use of "model" in scientific practice—i.e., all models are set theoretic structures. I will not be using "model" in this sense, and in that I will diverge at the outset from some (like Paul (2012)) who have discussed modelling in philosophy.[7]

The way I use the term "model" is broadly consistent with a philosophy of science tradition that includes Cartwright (1983, 1989) and Giere (1988, 2004), as well as many of those cited below. If you typically think of models as set-theoretic structures you will need to take this section as stipulating a new meaning for that term.

So what is a model? As the term covers many different scientific objects put to many uses, I will start with two more examples. Some models are material objects, like the molecular-structure models used by chemistry students. Modelling kits, such as the MolyMod system, come with coloured balls representing elements (red for Hydrogen, black for Oxygen), and grey connecting rods representing chemical bonds (short and stiff for single-valence, long and bendy for double-valence). With these kits, students build models of simple molecules like H_2O, and more complex polymers like PVC. We

[6]This section draws heavily from (Roussos, 2020, unpublished).

[7]This is not to deny that some models are set theoretic structures, but rather to deny the reductive definition for all models.

call the real-world system under study the target, and the plastic object the model. The model of H_2O involves one red ball connected by two short grey rods to two smaller black balls, in a wide V shape. The student learns about the structure of the molecule, H_2O, by examining the plastic model.

More commonly, models are theoretical rather than material. My model of the fish population is a good example. What is the model in this case? It is described by a series of written statements (like those above), often accompanied by equations (e.g., $N_{t+1} = 4N_t(1 - N_t)$), and perhaps illustrated with a diagram. But the system we are investigating when we use the model is not identical with any, or all, of these physically instantiated parts; it is what those descriptive elements specify (Mäki, 2009, p. 33). There are several philosophical accounts of what such "theoretical models" are, but for now we need only note that, whatever they are, they are not material objects.

Finally, some models have no target. Architectural plans for buildings which will never be built are models, as are theoretical models for ether or phlogiston, substances which do not exist. So, a definition of models must rely neither on a concrete model system, nor on a concrete target system.

Philosophers of science have developed a rich literature on the representational function of models, their ontology, epistemology, and implications for scientific realism (see Frigg and Hartmann, 2018). I will here draw attention to a few lessons learned in this literature, for comparison with the practice of moral philosophy. [8]

(1) Modelling is characterised by indirect or proxy inquiry (Giere, 2004; Godfrey-Smith, 2007; Weisberg, 2007b). Instead of studying the natural system (e.g., by counting fish), modellers describe and investigate a "model system". The model system is taken to (partially) represent the target natural system. Modellers then infer facts or generate hypotheses about the target system based on their investigation of the model system.[9]

(2) Models present an idealised and distorted picture of the target system (Frigg and Hartmann, 2018; Weisberg, 2007a). Many real-world systems cannot be investigated directly, due to incomplete theories or severe computational complexity. To make progress, scientists simplify the system under investigation, by changing or leaving out aspects of the real system. They work to identify the features of the system most salient to their investigation (Weisberg, 2013, p. 4). The frictionless plane is a classic example: no real surface is frictionless, but it is fruitful to take a surface to be frictionless when investigating inertial motion of objects on an inclined plane.

There is an extensive literature in idealisation in science; I will note two distinctions drawn in that literature for use here. There are different kinds of idealisations: Galilean and Aristotelian (Frigg and Hartmann, 2018). Galilean idealisations introduce deliberate distortions to some properties of the system under investigation. For example, the friction of the plane is deliberately changed in the representation. Aristotelian

[8] While they are not without opposition, I aim to use relatively "mainstream" views in the philosophy of modelling.

[9] In cases where the model is target-less, they may still be taken to be representational in a weaker sense that I won't discuss here (but see Frigg and Nguyen, 2016).

idealisations leave out features of the system that are not relevant to the problem being studied, to allow us to focus on or isolate a limited set of properties. For example, my population growth model considers only the rate of reproduction of fish and leaves out all their other properties.[10]

There are also different motivations for idealisations (Musgrave, 1981). A modeller might take a property to be negligible, believing that for the purposes of the current investigation it will make no difference to distort/exclude it. For example, we might consider falling objects and idealise by assuming there is no air resistance because we believe it to be of negligible importance. Another way of putting this is that the idealisation functions well when it is true that the effect of air resistance is small, so that the model's claim that air resistance is zero is approximately true.[11] Alternatively, the modeller might know that the property is not negligible in all cases but want to model only those cases where it is so. Musgrave calls this a domain idealisation: it justifies itself "automatically" by restricting the class of cases the model applies to. Finally, the modeller might think that there are no cases where the property is negligible but distort/exclude it anyway because its presence in the model makes things too complex to handle. Musgrave calls this a heuristic idealisation, and presents it as part of a process of inquiry: we simplify the model by setting air resistance to zero now, with the hope that once we have established the model we can factor air resistance in later. Note that negligibility, domain-restriction and heuristic necessity are species of justification—the same idealisation can be justified in each way, depending on the modeller and the circumstances.

(3) Models are built for a purpose, and so perform well only within a restricted domain of applicability (Parker, 2009; Teller, 2001; Weisberg, 2007b). "Purpose" consists of the main objects you're studying (e.g., ants rather than bears) and what you're trying to do (e.g., study group coordination). Prediction and explanation are common aspects of a model's purpose. The purpose establishes the basic domain of the model (it is an explanatory model of ant coordination). As Wimsatt (2007, p. 15) points out, models are often used to isolate particular mechanisms or concepts for study. This purpose motivates the idealising assumptions, which may further restrict the domain of applicability as discussed above. I'll refer to the combination of purpose and domain as the model's scope.

The purpose-driven nature of modelling means that model-based sciences often contain multiple, disagreeing models of the same phenomena. 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 models the liquid as a continuous incompressible medium. The second is interested in explaining diffusion, say of a drop of ink in water. It models 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, p. 401). Each is highly successful at its purpose, i.e., prediction of the

[10]Some authors call Aristotelian idealisations "abstractions," though usage is by no means uniform.

[11]I don't want to be committed to an approximate truth account of idealisation here; I am merely presenting some ways idealisations are thought of.

relevant kind of behaviour. Moreover, 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.

5 The methodology of modelling

The foregoing characteristics of modelling and models lead to certain methodological constraints for this kind of science. Idealisation is the lifeblood of modelling, but while it helps scientists make progress in investigations of complex systems, it introduces limitations. As Levins (1966) put it, modelling involves an inherent three-way trade-off between precision, realism and generality of scope.

> The multiplicity of models is imposed by the contradictory demands of a complex, heterogeneous nature and a mind that can only cope with a few variables at the time; by the contradictory desiderata of generality, realism, and precision; by the need to understand and also to control; even by the opposing aesthetic standards which emphasise the stark simplicity and power of a general theorem as against the richness and the diversity of living nature. These conflicts are irreconcilable. […] But the conflict is about method, not nature, for the individual models, while they are essential for understanding reality, should not be confused with that reality itself. (Levins, 1966, p. 66)

Levins goes on to claim that any model can at best maximise two of the three virtues (precision, realism, and generality), and notes that each position in the resulting triangular space of possibilities represents a strategy that may be best given a set of resources and purposes. This is offered both as an explanation of the social fact that there are multiple idealised models in some sciences, and as a justification of that practice (Weisberg, 2013, pp. 103–4).

This trade-off is thought to prevent scientists from developing a single "best" model for a complex system (Levins, 1966; Teller, 2001; Weisberg, 2013, Ch. 9). The resulting prevalence of multiple models of a single system also has methodological implications—most straightforwardly, we cannot take disagreements between, e.g., Teller's two models of water as a sign that one of them must be rejected. Each can be useful for its purpose.

On the realism front: 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. If there is an underlying fundamental theory (as is often the case in physics), this can help to identify artefacts. Another method for identifying such effects is sensitivity analysis.[12] This is a method for studying the uncertainty of a model, and allocating it

[12]Also called stability analysis, it is closely related to what Weisberg calls "parameter robustness"

to the sources of uncertainty in its inputs. In the use I am considering here, it involves varying assumptions in order to determine the effect that these variations have on the results. For example, let us consider again an idealisation of no air resistance, justified by a negligibility assumption. If we have set the parameter representing air resistance in our model to $k = 0$, we might vary this by considering small but non-zero values of k (small relative to some natural scale determined by the problem). The aim is to ensure that the results we get don't depend sensitively on the air resistance being exactly zero, and simultaneously to test that the negligibility assumption (about the real system) holds in our model—i.e., that small values of k make only small changes to the results.

The result of this kind of investigation is what Frigg and Nguyen (2016) call a "key". By analogy with a map's key, this is a legend that tells the user how to interpret what they're seeing. It specifies how results from the model should be taken to relate to the world, covering issues of precision and realism: a key might specify that some precise number generated by the model should be taken as a prediction for the real system only to within some error-margin; or it might identify some element of the model as an artefact, not to be imputed to the target at all.

As the above implies, criticising models is a complex business. As models have restricted domains, and specific purposes, the most natural way to critique a model is by examining its performance of its purpose within its domain. Performing poorly on other tasks, or in other domains, does not necessarily count against a model. It can do so if two models are being compared, and the one performs better on the shared purpose and has wider scope (either wider domain or the ability to fulfil multiple purposes). Put another way, models are not sensitive to counterexamples the way that fully general accounts are. Saying "here is a case that isn't like your model predicts" matters only if the case is in scope. Similarly, saying "your model says things are like so, but here is a case where they aren't" only matters if that feature of system in the model is intended to be imputed to the target. If the model's key identifies the feature as an artefact or says it should be imputed in some modified form, then the disagreement between the model's properties and the target's properties is irrelevant.

Why use models, if this is the case? Because they are helpful tools, which have been successfully put to explanatory, predictive and exploratory ends. They have proved a crucial part of the progress of science. Sometimes that progress is very localised: we gain understanding of a particular fish population in a particular lake. (But this is understanding we lacked before!) Sometimes, as noted by Wimsatt, it is progress on the path to developing a more systematic and wide-reaching theory.

5.1 The relation of models to theories

As I began this essay with the "anti-theory" critique of moral philosophy, I ought to say something about theories in science and their relation to scientific models. Unfortunately, "theory" is a much disputed term in the philosphy of science, and so I can once

(Weisberg, 2013, p. 159).

again offer only partial illumination. It is often defined by example: Newton's theory of gravitation and Einstein's general relativity are exemplars.

Scientific theories are tightly associated with laws. Laws are 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, p. 75). Theories seek to explain that uniformity, offering explanations and understanding of the phenomena in question, and allowing for predictions via the laws. Theoretical laws involve the introduction of theoretical concepts, which go beyond what can directly be observed.

Following the development of modern logic by Frege and others, and the "linguistic turn" in analytic philosophy, philosophers began to analyse mathematical and scientific theories in formal languages (see Glymour, 1999, for a historical discussion). In one resulting tradition of philosophy of science, theories came to be understood as sets of sentences in such a formal language. These sets are consistent, deductively closed, and (sometimes) axiomatisable.[13] In this "syntactic view", models in science are taken to be identical to models in logic: structures which interpret the theories. These structures are "models *of* theories". This usage survives outside of the syntactic view, and is used in circumstances when the theory is prior to the model, and the model is constructed using (parts of) the theory. One could describe a model of the solar system which is built from Newtonian mechanics in this way. In the contrasting semantic view of theories, theories are *collections of models*. Models are the prior objects here, and again are typically taken to be set-theoretic structures (Suppes, 1969).

In the latter half of the 20th century, philosophy of science turned its attention to scientific *change*. With this shift in attention, the concept of theory widened or was embedded in a wider concept. Kuhn presents what is probably an extreme in terms of breadth with his concept of a paradigm (in which a theory is embedded). A paradigm includes "the entire constellation of beliefs, values, techniques, and so on shared by the members of a given community" (Kuhn, 1970, p. 175). Lakatos, too, sought to shift attention from the theory to the research programme, which might be characterised as a series of theories marked by a continuity of methodology (Lakatos, 1970).

Regardless of the precise meaning of theory, much has been said about the link between models and theories.

One influential idea is that models are instruments which mediate between theories and reality (Morgan and Morrison, 1999). In this view models are independent from theories. A model might be constructed from (parts of) different theories, as well as detailed knowledge about instruments, approximation schemes, and other tools that

[13] 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.

are not parts of any theory (Cartwright, 1999). They may be required for prediction or explanation, with theories being too abstract to do that work for any particular circumstance.

Frigg and Hartmann (2018) discuss several ways that models and theories may interact in the practice of science. Models may be a means to explore theory, or to complement one. These uses may occur when the theory is very complicated and difficult to apply in full. Or perhaps the theory is leaves open certain questions, which a model fills in for particular cases. Models can also make quantitative what was only qualitative in the theory.

Some models exist entirely without theory. Recognition of this in philosophy of science arose from the study of biology, which makes extensive use of such models, but they are present in many sciences (see Frigg and Hartmann, 2018, S4.2., and references therein).

Wimsatt (2007, p. 104) highlights that in some such cases, multiple idealised models can support the development of theories (conceived of as broad explanatory structures involving laws), through the examination of results on which all models agree. This is a particularly useful technique in situations without underlying pre-existing fundamental theory, such as some areas of biology (Weisberg, 2013, p. 156).

6 Sketches of modelling in moral philosophy

In this section I present a series of short, suggestive characterisations of bits of normative ethics as models.[14] I am not trying to give a precise characterisation of what counts as a model in normative ethics, because no such characterisation exists in science either. I see many different models in ethics, of different kinds. Some are normative models of agents, and operate in a way that is similar to scientific models in that they represent some parts of the natural world. Others are models of goodness, and operate more like abstract models in theoretical physics, representing something that isn't a concrete part of the natural world.[15]

The picture I have in mind is this. Like scientists, moral philosophers begin with a set of "data": observations of moral life, and our moral judgements. They discern certain patterns amongst these, which they investigate, seeking eventually to systematise them in a moral theory. There may be some empirical regularities (e.g., common judgements, apparent norms), which will be explained 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. Ethical theorising therefore proceeds by modelling: restricting oneself to a limited domain (such as distributive questions for social planners, or justice-as-fairness), making distorting

[14]This section is somewhat anthropological. Having observed and interacted with ethicists, I offer a characterisation of their practice as modelling. This is a characterisation that would be more naturally made by themselves, once modelling talk enters their methodological lexicon.

[15]Putting aside forms of metaethical naturalism.

idealisations in order to simplify and facilitate systematisation.

As in science, these models often make qualitative principles quantitative. They allow one to use a general principle, like a Kantian maxim or a statement of act utilitarianism, to make judgements about particular cases.

Some models are intended to be "predictive"—which I interpret here as meaning rendering the correct judgement about a case. 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.[16]

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 satisfactory; 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.

6.1 Theory and Anti-theory

I now suggest that the way to understand the theory/anti-theory debate is as a dispute about whether ethics should be in the business of building theories, where that term is taken from or at least closely analogous with scientific theory. 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.

It is no mistake that the concept of theory in ethics is roughly the same as the concept found in science, and sometimes explicitly identical with the concept explicated by philosophers of science. In Baier's view, moral philosophers since the Enlightenment have been increasingly taken with analogies between ethics on one hand, and law and the sciences on the other. Ethical laws lie somewhere between laws of nature found in scientific theories and laws enforced in the courts (Baier, 1989). In this sense my methodological project is antithetical to the anti-theory project of Baier. Where I draw further inspiration from scientific methodology, she writes: "Philosophy, these days, seems in its methods not the queen of the sciences, expecting others to listen to her, but the social mime, drawing her procedures as well as her economic support from other

[16]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.

sources" (Baier, 1989, p. 43).

As I am seeking to characterise the practice of normative ethics, I will offer two interpretations of this 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. This is not necessarily to say that the interpretations are equally good; it may turn out that ethicists are doing better on one interpretation than on the other. They also need not compete—the first interpretation might better fit some parts of ethical theorising, and the second, others.

First interpretation: We have ethical theories, or partial 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 seem often to be literally false. Skilful practitioners must learn how to use models, approximation techniques, and various instruments, to connect these theories with reality.

Second interpretation: We do not 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, modelling which hopes to develop a theory. What is 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 forcefully challenged the notion that, even there, we have such laws. In *How the Laws of Physics Lie*, Cartwright argued that many laws of nature are literally false—what they tell us is not what happens. Those which are true are ceteris paribus laws, and apply only under abstracted and idealised conditions that are rarely realised in nature (Cartwright, 1983). These laws can still be thought of as true, and they 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, removing the nuance, complexity, and difficulty of moral reasoning. We now see that theories in science are themselves highly abstract and distant from the empirical phenomena they purport to explain. This has been noted before, and indeed Williams responded that the crucial difference is that in science the theories answer to the truth, which allows for them to be successfully general despite their abstraction (Fotion, 2014, pp. 55–56). Ethics, by contrast, is local and ethical statements are not truth-apt.

Setting aside the metaethical question here, I think that what has been neglected is the role of models in connecting the abstraction of scientific theory with the messy reality of local observation. Much of what Williams objects to in utilitarianism is that

its operations don't reflect the operations that real agents would carry out: think of the one-thought-too-many objection, or the complaint discussed above that utilitarianism offers easy verdicts to difficult questions. The mistake here seems to me to be a misunderstanding of which parts of the theory (utilitarianism) or rather its model (a filling in of the details, and assignment of numbers), 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, or to correspond to the cognitive processes by which that verdict would be arrived at by an actual agent.

Next, consider the complaint that theories require definiteness of meaning while moral norms are in reality vague. Baier's argument that seeming laws like "don't kill" are woven in to 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. Clearly causation will not be central to the success of moral theorising, but I think that this methodological analogy illuminates where anti-theorists need to focus their criticism.

The second interpretation does better against the remaining anti-theory arguments, I feel. On this interpretation we have no theory, and what we call a theory (e.g., utilitarianism) is "just a model".

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 is a form of idealisation, perhaps an Aristotelian abstraction or leaving out. It may be justified heuristically, 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. Finally, one might take the anti-theorists' insistent focus on moral conflicts as a prompt to study such dilemmas in a model, in order to illuminate their characteristics.

This no-theory interpretation can also answer Baier's definiteness worry. Here I would focus on the claim 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. But the proof of the pudding is in the eating, and so the modeller must generate some useful conclusions from this analysis. But there is nothing objectionable about its mere presence in moral philosophy, so long as we don't mistake

our models for theories.

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 modelling response. It is one of the distinctive features of modelling that we find a proliferation of models which 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![17]

There are three parts to the modeller's reply. The first is simply to insist that we are working in a complex, contradictory domain. *We do not have theories*, in the strong sense discussed above, though we are engaged in theorising. 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. Some contrary piece of data only *fails* to be a counterexample if it is genuinely outside of the model's purpose. "It is just a model" cannot be a Get Out of Jail Free card, it is a description of a careful and principled methodological approach. Below I give the example of regarding prioritarianism as restricted to the domain of distributive policy questions. 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—having wider scope is a virtue.

Moral particularism. The foregoing discussion also gives us a way to think about

[17]Acknowledgement.

moral particularism. Particularism is something like the view that not only is ethical theory impossible, but there is no middle-ground whatsoever. We must confront particular cases in all their granularity, rather than attempt any systematisation (e.g., Dancy, 2017). Using Levins' three-way trade-off framework, we can understand particularists as philosophers who eschew generality, and instead optimise for realism. [18] But rather than viewing this as one methodology within a field of potential strategies for inquiry, as the modeller does, particularists take generality to be a bad goal for moral philosophy.

Modelling offers 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, p. 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:

> A feature can make one moral difference in one case, and a different difference in another. Features have, as we might put it, variable relevance. Whether a feature is relevant or not in a new case, and if so what exact role it is playing there (the "form" that its relevance takes there) will be sensitive to other features of the case. This claim emerges as the consequence of the core particularist doctrine, which we can call the holism of reasons. This is the doctrine that what is a reason in one case may be no reason at all in another, or even a reason on the other side. (Dancy, 2017, S3)

This brings us back to Cartwright on laws of nature. Recall that to Cartwright, laws are literally false in much the way that particularists claims that the maxims in ethical theories are false. Cartwright argued 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

[18] I am not sure if particularists have a characteristic stance on precision.

> 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, as I said above, won't do for ethics. But here is what I want to take away from this: under our analogy, the reasons holist denies that morality has anything analogous to nature's capacities. This clarifies what the debate is about. The modeller, or aspirant theorist, may wish to reply that morality does involve the action of a constant tendency or capacity, but that they are only observed in the special circumstances described by the ceteris paribus clause. The mere fact that moral laws don't straightforwardly apply in observed cases is not an argument in favour of particularism. (What exactly the moral equivalent of "capacities" are I leave to moral philosophers.)

Having sketched this high-level picture of modelling in ethics, I will not turn to two parts of normative ethics which seem to me to display particular features that are best understood as modelling.

6.2 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 concerned with 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 these paradoxes and associated impossibility results arise. But they are not to be regarded as merely a problem for welfarists, claims Gustaf Arrhenius:

> It might be tempting for those who have little sympathy with utilitarian thought to try to set the problems raised by the above paradox to the side, thinking that it is a problem only for utilitarians or for those committed to welfarism [...] However, 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, this 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—to be precise, it is a form of Aristotelian idealisation in which it is assumed that these other factors can be left out of the model entirely, 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, but they also play a fairly explicit role in the practice of many sciences, most notably economics.

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, p. 304)

This seems to me to clearly be a model.[19]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) developing this mathematical representation requires making structural choices, in representing things one way rather than another, and (2) the idealisations involved in so doing.

[19]Though we needn't go into the details, here is a brief explanation: A welfare level (e.g., W_u) is a set containing populations (e.g., A) of equal welfare, where population itself a set of lives. The number

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 in fine-grained, in the following sense (Arrhenius, forthcoming, p. 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, p. 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). But here I would make 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. Now, Cartwright's is of course not the only game in town. But however one explicates it, modellers must engage in careful work to get their results to apply in messy real situations, either de-idealising the model where possible, 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.

---

of lives in a population is denoted $N(A)$. A range of welfare (e.g., $R(u, v)$) is a collection of ordered welfare levels, represented by its top and bottom points. So $R(u, v)$ is the set of welfare levels starting with W_u, the lowest ranked level in the set, and ending with W_y, the highest. The special symbol "1" is reserved for the welfare level that is just above the level at which life is not worth living.

	The first child	The second child
City	20	10
Suburb	25	9

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

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.[20]

6.3 Distributive theory

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. Here we face the same choice as in section 6.1, 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 distribute 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, pp. 81–83)

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

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

> 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, p. 83)

Here we see a clear example of modelling. The vignette contains no numbers, they are introduced as a thinking aid, along with some particular interpretative principles. Importantly, we are told which features to disregard—the precision is an artefact that the modeller, Parfit, wishes us not to impute to the target system. This structure allows Parfit to create models of the two views, egalitarianism and prioritarianism, in order to investigate their properties in a precise way. For example, egalitarianism, a view about people being equally well off, becomes a view about equality between numbers representing people's welfare. (Once again, we can regard these as models of theories, or models in the absence of full theories.)

Taking up the same case in a recent discussion of prioritarianism and egalitarianism, Otsuka and Voorhoeve (2018) introduce some additional structure that it is useful to reflect on. In their case, there is uncertainty about the outcome (in the form of objective, given probabilities for them). 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; 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.

Modelling comes with methodological constraints such as those discussed above for science. The choices made in their construction will have consequences for each model's balance of realism, generality and precision. These idealisations and representational choices will constrain the inferences one can draw from them. Depending on how the idealisations are motivated, they might restrict the domain to which the model successfully applies. Alternatively, the modeller will need to think carefully about how the results depend on the idealisations—perhaps aided by sensitivity analysis.

What might it mean to say that a model of this sort has a restricted domain of application? Here is one example of how one might apply the idea of domain-restriction to prioritarianism.[21] One might think that 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. Thus, the model deliberately neglects considerations that would introduce partiality. This means that the structure of the preferences of the ideal prioritarian is such that the model is simply unsuited to applications to friendship, or romantic relationships. So on this view, for example, using prioritarianism to think from a first-personal perspective about whom I ought to marry (as Arneson, 2010, does) is simply a mistake.

The flipside of this restriction is immunity to counterexamples that are outside of the domain. If one regards prioritarianism as a model with this domain, then its performance as an ethics of marriage, or friendship, is simply irrelevant. An interlocutor might say to the prioritarian that their view is implausible because it would lead us to obsess over small differences between friends, or to cast ourselves as victims in the face of our friends' successes. The social policy prioritarian needn't even answer these challenges, for they are outside of the intended domain of the model.

This will likely feel like cheating to the ethicist who has made a career of thinking of counterexamples. I refer them back to my three-point reply to the child-torturing case

[21] Acknowledgement.

above. It will surely take time to unlearn the habits of a lifetime, but there is nothing illegitimate about the methodology of modelling.

A final constraint: these idealisations and restrictions limit the applicability of model results. Otsuka and Voorhoeve are admirably precise theorists, who are careful to clarify which assumptions they are using. But as I read their work, there is insufficient attention paid to how their results may depend on these assumptions and thus be limited by them. Real agents do not satisfy the VNM axioms, nor are they required to on many theories of rationality.[22] Does this difference matter for the results in Otsuka and Voorhoeve's paper? We are not told, though Otsuka (2015) has taken up the question elsewhere. But as this is a model, this should be front and centre—it is crucial to even understanding the results! Note that if the results *are* restricted in their generality, this does not make them worthless. Models are part of a highly successful strategy of inquiry in science. Making progress in a limited domain is still making progress.

7 Why use modelling in moral philosophy?

Having shown that moral philosophy can and does use models, I conclude this discussion with a reflection on why we would want to use them. As you might expect, the benefits are much the same as those of theories. Modelling comes with some limitations, but it is possible when we do not have theories.

This indeed is the first benefit. If one takes seriously the challenge that impossibility theorems in population ethics pose to moral theory, then one might be tempted to despair: there is no satisfactory axiology! But this is where we can take heart from the success of modelling in scientific domains without known fully general and consistent theories. Rather than seeking principles which accord with all of our intuitions, modelling requires that we seek plausible restrictions to which intuitions we seek to satisfy—not in the sense of rejecting an intuition outright, as in the traditional story about achieving reflective equilibrium, but locally, for a specified purpose.

Modelling need not conflict with the goal of theory development; as we have seen, in the sciences it supports it. Additionally, modelling can go with the use of cases, as Parfit uses it, or could be used in their absence by critics of the "method of cases" (e.g., Machery, 2017). Modelling is also not a competitor to the well-established goal of reflective equilibrium. I take this to be a process of testing principles against judgements, and iteratively revising both until equilibrium is reached. But this process needn't occur through the brute testing of judgement and principle against cases. Modelling can support it, by providing new methods for testing out the implications of various sets of judgements and principles, and precisely identifying their consequences in particular cases.

Models help to clarify and discipline one's thoughts. By making definitions and

[22]For example, VNM agents are strictly risk neutral, which seems to me to have no ethical or rational motivation.

claims precise, one can test their implications and interactions. For example, Nebel and Stefánsson (unpublished) use a model to show how a seemingly plausible principle, aversion to inequalities when the stakes are small, commits prioritarians to aversion to situations in which half a population gains a very large quantity of well-being. Put precisely, they claim prioritarians are committed to preferences under which "a distribution in which 2n people are at any level w is preferred to one in which n people are at $w - 8$ and n are at $w + G$, *no matter how large G is*" (Nebel and Stefánsson, unpublished, p. 11). Note that the modelling exercise doesn't replace the work of moral philosophy: one must still determine whether the model's result is a bad thing for the prioritarian. What the model achieves is clarity, ease of inference, and a demonstration of why the result follows from the seemingly plausible principle.

Recognising that we are modelling, and adopting modelling more explicitly as a methodology, will bring changes to the practice of moral philosophy. Probably the most significant methodological upshot is a re-evaluation of 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. If my characterisation of current work in normative ethics is correct, much of this practice is misguided. If my methodological recommendations are taken up, new ways of working will need to be developed. Rather than being trained to seek counterexamples, graduate students in ethics will need to be trained in modelling: the careful use of idealisation, the analysis of the effects of idealisations through sensitivity analysis, and a new comfort with making progress locally, in the absence of theory.

References

Anscombe, G. E. M. (Jan. 1958). "Modern Moral Philosophy". In: *Philosophy* 33.124, pp. 1–19. DOI: 10.1017/S0031819100037943.

Arneson, Richard J. (2010). "Democratic Equality and Relating as Equals". In: *Canadian Journal of Philosophy Supplementary Volume* 36, pp. 25–52. ISSN: 0229-7051, 2633-0490. DOI: 10.1080/00455091.2010.10717653. URL: https://www.cambridge.org/core/product/identifier/S0229705100000458/type/journal_article (visited on 10/09/2020).

Arrhenius, Gustaf (Oct. 2000). "An Impossibility Theorem for Welfarist Axiologies". In: *Economics and Philosophy* 16.2, pp. 247–266. ISSN: 0266-2671, 1474-0028. DOI: 10.1017/S0266267100000249. URL: https://www.cambridge.org/core/product/identifier/S0266267100000249/type/journal_article (visited on 09/26/2020).

— (forthcoming). *Population Ethics*. Oxford University Press.

Baier, Annete (1989). "Doing Without Moral Theory". In: *Anti-Theory in Ethics and Moral Conservatism*. Ed. by Stanley G. Clarke and Evan Simpson. Albany: State University of New York, pp. 29–49.

Baier, Annette (1985). *Postures of the Mind*. Minneapolis: University of Minnesota Press.

Bovens, Luc and Stephan Hartmann (2003). *Bayesian Epistemology*. Oxford: Oxford University Press.

Budolfson, Mark and Dean Spears (forthcoming). "Does the Repugnant Conclusion Have Important Implications for Axiology or for Public Policy?" In: *Oxford Handbook of Population Ethics*. Ed. by Tim Campbell, Krister Bykvist, and Gustaf Arrhenius. Oxford: Oxford University Press.

Cartwright, Nancy (1983). *How the Laws of Physics Lie*. Oxford University Press. ISBN: 978-0-19-159715-2.

— (1989). *Nature's Capacities and Their Measurement*. Oxford University Press.

— (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Chappell, Sophie-Grace (unpublished). "If Not Moral Theory, Then What?" In: *Epiphanies*. manuscript, pp. 1–46.

Clarke, Stanley G. (1987). "Anti-Theory in Ethics". In: *American Philosophical Quarterly* 24.3, pp. 237–244.

Colyvan, Mark, Damian Cox, and Katie Steele (2010). "Modelling the Moral Dimension of Decisions". In: *Noûs* 44.3, pp. 503–529. ISSN: 0029-4624. JSTOR: 40959656.

Dancy, Jonathan (2017). "Moral Particularism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2017. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2017/entries/moral-particularism/ (visited on 09/26/2020).

Eva, Benjamin and Stephan Hartmann (2019). "On the Origins of Old Evidence". In: *Australasian Journal of Philosophy* forthcoming in print, pp. 1–14.

Fotion, Nick (2014). *Theory vs. Anti-Theory in Ethics: A Misconceived Conflict*. Oxford University Press. ISBN: 978-0-19-937354-3.

Frigg, Roman and Stephan Hartmann (2018). "Models in Science". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2018. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2018/entries/models-science/ (visited on 07/04/2018).

Frigg, Roman and James Nguyen (2016). "The Fiction View of Models Reloaded". In: *Monist* 99.3, pp. 225–242. DOI: 10.1093/monist/onw002.

Giere, Ronald N (1988). *Explaining Science*. Science and Its Conceptual Foundations. Chicago: University of Chicago Press. URL: http://www.press.uchicago.edu/ucp/books/book/chicago/E/bo3622319.html (visited on 07/16/2018).

— (2004). "How Models Are Used to Represent Reality". In: *Philosophy of Science* 71, pp. 742–52.

Glymour, Clark N. (1999). "Realism and the Nature of Theories". In: *Introduction to the Philosophy of Science*. Ed. by Merrilee H. Salmon. Indianapolis and Cambridge: Hackett.

Godfrey-Smith, Peter (Feb. 15, 2007). "The Strategy of Model-Based Science". In: *Biology & Philosophy* 21.5, pp. 725–740. ISSN: 0169-3867, 1572-8404. DOI: 10.1007/

s10539-006-9054-6. URL: http://link.springer.com/10.1007/s10539-006-9054-6 (visited on 07/04/2018).

Hempel, Carl G. (1966). *Philosophy of Natural Science*. Foundations of Philosophy. London: Prentice Hall.

Kuhn, Thomas S. (1970). *The Structure of Scientific Revolutions*. Second. Chicago: University of Chicago Press.

Lakatos, Imre (1970). "Falsification and the Methodology of Scientific Research Programs". In: *Criticism and the Growth of Knowledge*. Ed. by Imre Lakatos and Alan Musgrave. Cambridge: Cambridge University Press, pp. 91–196.

Leitgeb, Hannes (2013). "Scientific Philosophy, Mathematical Philosophy, and All That". In: *Metaphilosophy* 44.3, pp. 267–75.

Levins, Richard (1966). "The Strategy of Model Building in Population Biology". In: *American Scientist* 54.4, pp. 421–431. ISSN: 0003-0996. JSTOR: 27836590.

List, Christian and Laura Valentini (2016). "The Methodology of Political Theory". In: *Oxford Handbook of Philosophical Methodology*. Oxford: Oxford University Press. URL: http://personal.lse.ac.uk/list/PDF-files/MethodologyPoliticalTheory.pdf.

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

Machery, Eduoard (2017). *Philosophy Within Its Proper Bounds*. Oxford: Oxford University Press.

McCarthy, David, Kalle Mikkola, and Teruji Thomas (Nov. 3, 2019). "Aggregation for Potentially Infinite Populations without Continuity or Completeness". In: arXiv: 1911.00872 [econ]. URL: http://arxiv.org/abs/1911.00872 (visited on 09/24/2020).

McKeever, Sean and Michael Ridge (2015). *Obvious Objections*. Oxford University Press. ISBN: 978-0-19-178165-0.

Mäki, Uskali (2009). "MISSing the World. Models as Isolations and Credible Surrogate Systems". In: *Erkenntnis* 70, pp. 29–43.

Morgan, Mary S. and Margaret Morrison, eds. (1999). *Models as Mediators*. Cambridge: Cambridge University Press.

Musgrave, Alan (1981). "'Unreal Assumptions' in Economic Theory: The F-Twist Untwisted". In: *Kyklos* 34.3, pp. 377–87.

Nebel, Jacob M. and H. Orri Stefánsson (unpublished). "Calibration Dilemmas in the Ethics of Distribution".

Nozick, Robert (1981). *Philosophical Explanations*. Oxford: Clarendon Press.

Nussbaum, Martha (2000). "Why Practice Needs Ethical Theory". In: *Moral Particularism*. Ed. by Brad Hooker and Margaret Olivia Little. Oxford: Clarendon Press, pp. 234–345.

Otsuka, Michael (2015). "Prioritarianism and the Measure of Utility". In: *Journal of Political Philosophy* 23, pp. 1–22.

Otsuka, Michael and Alex Voorhoeve (2018). "Equality versus Priority". In: *Oxford Handbook of Distributive Justice*. Oxford: Oxford University Press. URL: http:

//personal.lse.ac.uk/OTSUKAM/M%20Otsuka%20%20A%20Voorhoeve%20(final)%20w%20author%20edits%202015-06-15%20REVISED%20clean.pdf.

Parfit, Derek (1984). *Reasons and Persons*. Oxford: Clarendon Press. ISBN: 0-19-824908-X.

— (2002). "Equality or Priority?" In: *The Ideal of Equality*. Ed. by Matthew Clayton and Andrew Williams. New York: Palgrave Macmillan, pp. 81–125.

Parker, Wendy S. (2009). "Confirmation and Adequacy-for-Purpose in Climate Modelling". In: *Proceedings of the Aristotelian Society, Supplementary Volumes* 8. JSTOR: 20619137.

Paul, L.A. (2012). "Metaphysics as Modeling: The Handmaiden's Tale". In: *Philosophical Studies* 160.1, pp. 1–29. JSTOR: 23262471.

Raphael, D.D. (1974). "The Standard of Morals: The Presidential Address". In: *Proceedings of the Aristotelian Society* 75, pp. 1–12.

Reutlinger, Alexander et al. (2019). "Ceteris Paribus Laws". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2019. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2019/entries/ceteris-paribus/ (visited on 09/26/2020).

Roussos, Joe (May 2020). "Policymaking under Scientific Uncertainty". PhD Thesis. London: London School of Economics and Political Science. URL: http://etheses.lse.ac.uk/4158/.

— (unpublished). "Formal Epistemology as Modelling". In: *manuscript*.

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

Teller, Paul (2001). "Twilight of the Perfect Model Model". In: *Erkenntnis* 55.3, pp. 393–415. JSTOR: 20013097.

Timmons, Mark (2012). *Moral Theory: An Introduction*. Lanham, MD: Rowman & Littlefield Publishers. ISBN: 978-0-7425-6493-0.

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

Weisberg, Michael (2007a). "Three Kinds of Idealization". In: *The Journal of Philosophy* 104.12, pp. 639–659. ISSN: 0022-362X. JSTOR: 20620065.

— (2007b). "Who Is a Modeler?" In: *The British Journal for the Philosophy of Science* 58.2. JSTOR: 30115224.

— (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

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

— (1981). *Moral Luck: Philosophical Papers 1973–1980*. Cambridge: Cambridge University Press. ISBN: 978-0-521-24372-8. DOI: 10.1017/CBO9781139165860.

Williamson, Timothy (2006). "Must Do Better". In: *Truth and Realism*. Ed. by Patrick Greenough and Michael P. Lynch. Oxford : New York: Clarendon Press ; Oxford University Press, pp. 177–188. ISBN: 978-0-19-928888-5 978-0-19-928887-8.

— (2017). "Model-Building in Philosophy". In: *Philosophy's Future: The Problem of Philosophical Progress*. Ed. by Russell Blackford and Damien Broderick. Oxford: Wiley.

Wimsatt, William C. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.