Normative Formal Epistemology as Modelling

Joe Roussos

November 2021

This is a preprint of a paper forthcoming in The British Journal for the Philosophy of Science. Please cite published version.

Abstract

I argue that normative formal epistemology (NFE) is best understood as modelling, in the sense that this is the reconstruction of its methodology on which NFE is doing best. I focus on Bayesianism and show that it has the characteristics of modelling. But modelling is a scientific enterprise, while NFE is normative. I thus develop an account of normative models on which they are idealised representations put to normative purposes. Normative assumptions, such as the transitivity of comparative credence, are characterised as modelling idealisations motivated normatively. I then survey the landscape of methodological options: what might formal epistemologists be up to? I argue the choice is essentially binary: modelling or theorising. If NFE is theorising it is doing very poorly: generating false claims with no clear methodology for separating out what is to be taken seriously. Modelling, by contrast, is a successful methodology precisely suited to the management of useful falsehoods. Regarding NFE as modelling is not costless, however. First, our normative inferences are less direct and are muddied by the presence of descriptive idealisations. Second, our models are purpose-specific and limited in their scope. I close with suggestions for how to adapt our practice.

1 Introduction

I will argue that much normative work in formal epistemology is best understood as a kind of modelling, where by this I mean the widespread and successful strategy of scientific inquiry.[1]

What is a model, and what is this method? I shall spend some time on this, but here is a brief example. In studying 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 ignore many features of the real pond and fish, such as the natural variation in fish size and reproduction. I treat time as discrete, and count only weeks. I ignore certain factors which I know 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. Crucially, I go on with my inquiry by studying this formal apparatus as a proxy for the pond itself. These are characteristic features of modelling as a method.

Talk about "modelling" as a method of philosophical inquiry is increasingly prevalent across a range of philosophical sub-fields. Williamson (2006, 186-7) defends modelling as a tool for developing clear arguments, and as a major source of philosophical progress (Williamson, 2017, 8). Leitgeb (2013, 273) cites modelling as a method for building inductive strength in an argument. Godfrey-Smith (2006, 2012) and Paul (2012) have discussed modelling as a practice in metaphysics. Titelbaum (2012) described his book as providing a framework for building models in formal epistemology, and has recently written on the nature of normative models (Titelbaum, b). Colyvan (2013) has written about idealisations in normative models, and Yap (2014) has characterised epistemic logic as a descriptive but idealised model.

My focus in this essay will be a group of philosophical approaches falling under the heading of "Bayesianism" in formal epistemology and decision theory: philosophical work which talks about agents having and updating probabilistic beliefs. I am interested in the use of Bayesian tools for normative ends: generating evaluative and prescriptive claims about how agents ought to be or what they ought to do, in order to be rational. Though my focus is on Bayesianism, it is intended to stand in for a wider range of philosophical work which makes use of idealised logico-mathematical representations of agents and their attitudes for normative ends; including other bits of epistemology (epistemic logic, belief revision theory, Dempster-Shafer theory, etc.), as well as work that goes under the

[1] I suspect that little of what I argue for here is deeply attached to mathematics. I begin with formal philosophy because it is a relatively clear case. But the characteristics which I argue make NFE modelling can also be found in some non-formal work. For this reason, the precise delineation between formal and non-formal epistemology is not of major concern to me. Some non-formal normative epistemology could be construed as modelling, and I think that modelling is a fruitful method for non-formal philosophers. I will not defend these claims here, but for a discussion of modelling in ethics, not dependent on formalism, see (Roussos, ms).

headings of social choice and opinion pooling. For this reason I will refer to my target as "normative formal epistemology" or NFE, without meaning to imply by my choice of examples that all such formal work in epistemology is Bayesian, nor that I can neatly delineate formal from non-formal epistemology.

This is a philosophy of science paper, where the target "science" is (this bit of) philosophy.[2] I analyse how philosophers are working and argue that much of this work is best understood as modelling. This may sound like I will describe empirical research into how philosophers in fact work, but that is not my intention. I intend to offer a rational reconstruction of the method of NFE, in the sense of Reichenbach (1938). Such a reconstruction has a quasi-descriptive part and a critical part.

I begin by showing that NFE can be described as modelling, by offering a reconstruction which has characteristic features of scientific modelling. In so doing, I provide an account of how normative work can be modelling, which is typically descriptive. I then offer philosophers a dilemma: modelling or bust. I contrast modelling with direct theorising, and argue that if we do not regard NFE as modelling, then it appears to be rife with falsehoods—itself a disqualifying characteristic for theory—and without a clear, justified method of sorting claims that are to be taken seriously from those that are to be ignored. In it in this sense that normative formal epistemology is "best" understood as modelling: this reconstruction best justifies its methods. Thus, epistemologists ought to be modelling. Once we accept this, we can turn to tuning our modelling methods. I argue that much more needs to be done to study the interaction between our idealisations and our normative conclusions. I then argue that our immediate goals need revision: while we may desire universal and perhaps necessary truths, what models give us is something much more limited.

2 The Target

In the kind of work I have in mind, philosophers typically start with an initial question/problem framed in natural language. Some principles of rationality governing the agents involved are chosen for investigation. These are translated into a formal language capable of representing agents, propositions, beliefs, and so on (as necessary). Constructing this formal apparatus often requires introducing additional structure, that is not motivated by the initial question but is internal to the process of representing it mathematically. The formal setup is then studied, and conclusions are drawn. Finally, these formal results are translated into conclusions about partial belief or decision-making.[3]

This sort of thing is now very common and so I assume quite familiar, but here are two examples for specificity. First, from Pettigrew (2016, 2):

> Let us suppose that Yasho has opinions about only two propositions:
>
> A: Sonya is a political activist and an accountant.
>
> B: Sonya is an accountant.

[2] If, like Morgenbesser (1967, xvi), you think that "philosophy of science is epistemology with scientific examples" then you can take this as an epistemology paper with formal epistemology examples.

[3] Formal epistemologists and decision theorists use many terms to refer to the belief-like attitude they study: partial belief, degree of belief, credence, comparative confidence, and no doubt others. At times these are used with subtle differences, but in this essay they should be taken to be synonymous.

> We represent his doxastic state by a function c that takes each proposition about which he has an opinion and returns a real number that measures his credence (or degree of belief) in that proposition. By convention, we take 0 to measure minimal credence, and 1 to measure maximal credence. Thus, we represent Yasho's doxastic state by a function $c : \{A, B\} \to [0, 1]$ (where $[0, 1]$ is the set of real numbers at least 0 and at most 1). We call c his credence function and $\{A, B\}$ his opinion set. In our example, Yasho has greater credence in A than in B: thus, $c(B) < c(A)$.

Second, another fairly representative passage from Earman (1992, 35-36) laying out a mathematical structure along with some interpretation.

> [S]ince probability is being interpreted as degree of belief, probabilities will be assigned to objects which express propositions, namely sentences.[. . .] Then a *probability function* Pr is a map from $\mathcal{A}$ to $\mathbb{R}$ satisfying at least the following restrictions:
>
> (A1) $\Pr(A) \geq 0$ for any $A \in \mathcal{A}$
>
> (A2) $\Pr(A) = 1$ if $\models A$
>
> (A3) $\Pr(A \vee B) = \Pr(A) + \Pr(B)$ if $\models \sim (A \& B)$
>
> Here $\models A$ means that A is valid in the sense that A is true in all [interpretations] or all possible worlds. [. . .] I assume at a minimum that $\mathcal{A}$ respects propositional logic. In this case (A1) to (A3) suffice to prove many of the familiar principles of probability, including the following:
>
> . . .
>
> (P2) $\Pr(A) = \Pr(B)$ if $A \leftrightarrow B$
>
> . . .
>
> (P4) $\Pr(A) \leq \Pr(B)$ if $A \models B$
>
> Here $A \models B$ means that A semantically implies B in the sense that B is true in every model or possible world in which A is true.

Note some obvious features of these passages. The agential attitude under study, degree of belief, is represented by a mathematical object. That object is a probability function and its mathematical properties are taken to represent features of the attitude. In Pettigrew's passage, the mathematical relationship $c(B) < c(A)$ represents Yasho having greater confidence that Sonya is a political activist and an accountant than that she is an accountant.

There are two roles played by the mathematics here. The first is description. Pettigrew is introducing some mathematics which represents some facts about Yasho. Earman begins at the more general level, introducing a framework for representing degrees of belief. Some features of this general framework match with patterns of facts about agents' actual degrees of belief. We are able to compare our strength of belief in different propositions and, when things go well, our degrees of belief are transitive and respect known logical relations between propositions. But this is not descriptive work, or not primarily descriptive. Bayesian epistemology is a theory of rational partial belief, and so the mathematics represents properties which agents ought to have. Indeed, the mathematics is used to generate normative claims like “if an agent has a certain degree of belief in A, and A implies B, then the agent ought to have at least as high a degree of belief in B as in A.”

My target is this kind of mathematical work which represents—in some sense—agents and their attitudes and which is deployed for normative ends.

One of my motivations is to find a way to understand the seeming falsehoods that this work involves. Here are two examples. Earman’s (A1) says that there is a probability value for every sentence in the set $\mathcal{A}$. Thus, the agent has an attitude to every proposition and, as these are represented by numbers, they are commensurable. We can compare the agent’s attitudes to any two propositions precisely and quantitatively: we can say how much more confident they are in one than the other. This quantitative comparability is critical for many decision theoretic applications of this framework. But, to some, the completeness and fine-grainedness of this comparability are neither normative nor true. How should we regard the fact that the mathematics seems to “say” that agents have complete and continuous credences, when in fact they don’t?

Earman’s (A2), (P2), and (P4), are part of the infamous “logical omniscience” of Bayesian agents. They say that agents (ought to) assign the maximum degree of belief in any sentence expressing a logical truth, that the same degrees of belief are assigned to logically equivalent sentences, and that degree of belief is monotonic over entailment. Agents who were really like this would have quite unusual beliefs! They would believe fully every truth of mathematics, believe equally in even opaque and complex but equivalent sentences, and their degrees of belief would track logical entailments—all of which is, clearly, far beyond us. Again, what should we make of this? Some argue that logical omniscience is a norm or ideal, of a sort (e.g., Lemmon and Henderson (1959), see Yap (2014) for further discussion of how omniscience is regarded in epistemic logic). But again, others disagree. For this latter camp, there is an important question of whether and when it matters that the framework has this false implication.

Below, I will argue that regarding NFE as modelling allows us to understand the nature of these claims, tells us why we don’t need to worry about them, and gives us a way to do so in a justified and systematic manner.

3 Normative Formal Epistemology *as* Modelling

In this section and the next I will show that the shoe fits. I will relate some lessons from the last five or six decades of philosophical study of scientific modelling, and then show that NFE has the characteristics philosophers have identified for models. This establishes the possibility and hopefully plausibility of interpreting NFE as modelling. I will then turn to arguing that we ought to do so.

A preliminary note: “Model” is used to mean many different things. I do not intend to use the meaning logicians give to the term, which is roughly an assignment of semantic values to a basic vocabulary, such that a specific set of sentences is rendered true. The way I use the term “model” is broadly consistent with a philosophy of science tradition that includes Cartwright (1983, 1989), Morgan and Morrison (1999) and the many others 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.[4]

[4] As Williamson (2017) points out, some models (in roughly my sense) in formal epistemology are also logic models, so my sense of model is wider than the logical sense. I am thus not denying that logic models play a role, they simply aren’t my focus.

We have already seen an example of a scientific model: the logistic growth model of my fish population, described by the second paragraph of this paper, and making use of the equation $N_{t+1} = 4N_t(1-N_t)$. Note the verbs I used there: the model is "described" by those sentences, and "uses" that equation. The model itself is a kind of abstract system which those elements specify (Mäki, 2009, 33). The model represents a target system in the world, in this case my fish pond. Some models have types of systems as their target, rather than particular systems—like Bohr's model of the hydrogen atom. Others appear to have no target, such as theoretical models for ether or phlogiston, substances which do not exist.

Here we have two important elements of modelling. First, modelling is characterised by indirect inquiry (Giere, 2004; Weisberg, 2007b; Godfrey-Smith, 2007). Instead of studying a natural system, modellers describe and investigate a "model system" which is the primary target of their investigation. Second, representation plays a crucial role in modelling: the model system is taken to (partially) represent the target natural system (Frigg and Hartmann, 2018; Frigg and Nguyen, 2020). Modellers infer facts or generate hypotheses about the target system based on their investigation of the model system. (In cases where the model is target-less, they are still thought to be representational in a sense to be discussed later.)

Indirect inquiry using representations is a neat description of much NFE. Consider Pettigrew's argument for Probabilism. "Probabilism is a coherence requirement. It says how a credence in a proposition should relate to credences in other, logically related propositions. It requires that an agent's credences obey the axioms of the probability calculus" (Pettigrew, 2016, 8). In his usage, "credence" refers both to the agent's attitude and to the function c which Pettigrew introduces to represent that attitude. I read the second sentence just quoted as referring to the attitudes: Probabilism is a norm which tells us how our degrees of belief should be related. It requires that the mathematical function which represents them obey the probability calculus—or, more carefully, that those attitudes be so representable. The bulk of Pettigrew's book is about these credence functions. He introduces various norms for beliefs, translates them into his mathematical framework, and then proves various results which build up to Probabilism. Credence functions are investigated as a proxy for studying credences, which they represent. In section 4 I discuss how the normativity of this work interacts with its use of representation.

The next characteristic feature of scientific modelling is its central use of idealisation. Rather than representing their targets with perfect accuracy, models present an abstracted and distorted picture of the target system (Weisberg, 2007a; Frigg and Hartmann, 2018). Many real-world systems cannot be investigated directly, due to incomplete theories, partial understanding, or severe computational complexity. Scientists work to identify the features of the system most salient to their investigation (Weisberg, 2013, 4). To make progress, they simplify the system under investigation, by changing or leaving out aspects of the real system—these changes are called "idealisations". 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.

The mathematical structures in NFE have many differences from the attitudes they represent. I have already mentioned the precise numerical comparability and completeness of partial belief, and logical omniscience. Logic-based models of belief often use sets of propositions which are closed under entailment, though the attitude represented, such as real belief or knowledge, is typically not so closed. Put in our new language: there are differences between the properties of the target of the model—the real agent and their beliefs—and the representation in the model. I propose that these are all idealisations, and I will discuss the role normativity plays in justifying some of them in section 4.4.

There are different kinds of idealisations (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 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, a population growth model considers only the rate of reproduction and leaves out all other properties.[5]

I suggest that the continuity of credence is a Galilean idealisation, as are the completeness and commensurability the Bayesian framework introduces. Thought of as a representation of an agent, Bayesian decision theory has many Aristotelian idealisations: the orthodox theory focuses on only two attitudes—belief and desire—leaving out other aspects of the agent's psychology.

Idealisations introduce falsehoods into the set of descriptions of the target system that the model generates. Some of these false descriptions are unexpected properties which emerge during the study of the model. These "artefacts" should not be used to generate predictions or explanations of the real system. Modelling requires skilful interpretation—modellers know that some descriptions generated by the model are not to be imputed to the target system, while others are. As a result, models require 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 in the model. It specifies whether and how results from the model should be taken to relate to the world. In section 6 I will argue that philosophical models, too, come with (often implicit) keys.

These falsehoods are not accidental, nor can they always be removed. They are often critical to generating the model's results, including true explanations of phenomena (Frigg and Nguyen, 2021). Models are crucial to the development of scientific understanding of the modelled system, despite their falsehoods (de Regt, 2017). This should temper some worries that describing NFE as modelling means giving up on true explanation or understanding as epistemic goals.

Idealisations, representational choices, and trade-offs in the modelling process are justified relative to the purpose of the inquiry. This purpose and the idealisations together fix a restricted domain of applicability for scientific models' results (Teller, 2001; Weisberg, 2007b). "Purpose" consists of what you're modelling (e.g., ants rather than bears) and what you're trying to do (e.g., study group coordination). This establishes the basic domain of the model (it is a model of ant coordination).

There are two important consequences of this. First, model-based sciences often contain multiple, disagreeing models of the same phenomena. Teller illustrates this with an example of two models of water. One is interested in the flow of water and wave propagation, and it models the liquid as a continuous incompressible medium. The other 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, 401). Neither should be thought to provide a definitive characterisation of water, and our understanding of water is enhanced by having both available. Second, philosophers have argued that the success condition for scientific models is not truth per se, but adequacy for purpose (notably Parker, 2009). This does not remove truth from the picture: it

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

remains the ultimate goal of inquiry. But models have different immediate aims, against which they are evaluated. Truth is approached more indirectly, through a series of model-based inquiries which delivery important insights about limited domains, or particular questions.

The goals of modelling will have to wait until section 6, but I will comment briefly on multiple models in NFE. Let us widen our focus out from Bayesianism to some other approaches I mentioned in the introduction: epistemic logic, belief revision theory, and Dempster-Shafer theory. There is, I think, a perfectly natural sense in which these approaches don't compete but rather are suited to different purposes. Bayesian methods don't foreground knowledge, and so epistemologists working on knowledge find tools more suited to their ends in epistemic logic. Because of the lottery and preface paradoxes, it is difficult to account simultaneously for dichotomous and graded belief. So, for questions framed naturally in terms of full belief, a model which neglects degrees of belief is often suitable. Epistemologists interested in how evidence supports beliefs might want to formally represent the difference between the evidential situation of an agent with a known fair coin, and that of an agent with a new coin of unknown bias. Dempster-Shafer belief functions are intended to measure support, and using them one can express explicitly the evidential differences between the risky and ambiguous coin (Genin and Huber, 2021).[6]

4 Normativity and Modelling

I hope that you now see enough of a rough fit between normative formal epistemology and modelling to be worth investigating. I will now show in detail how we can fit normativity into an account of modelling, while retaining the characteristics described above. This serves two purposes: it continues my reconstruction of NFE as modelling, and motivates for drawing on the philosophy of science literature in improving our practice in epistemology and decision theory. Along the way, I characterise what it means for something to be a normative model. My account differs from an existing account of normative models, due to Titelbaum (b), and so I discuss the differences between our views below.

Here is a summary of the various pieces of the account. In science, models often represent a target system which they are "models of". In NFE, these are models of agents, like you and me, and our attitudes like partial belief. In science, models are built to "fit" some data. In NFE, this data is a mixture of descriptive facts about agents' attitudes and what we might call "normative data" (following Titelbaum, b). In science, models employ idealisations—deliberate distortions or omissions of properties of the target system. In NFE, we employ both "descriptive idealisations" and "normative idealisations". In science, models are used to predict, explain, and understand the target system. So too in NFE, though in place of prediction we have something like the generation of normative claims, which are tested against the "data"—our considered judgements.

4.1 Target and Representation

It is natural to say that models in NFE are representations of beliefs, desires, and agents. The agents and attitudes are not represented in a descriptively accurate fashion, because this work is normative. Instead these representations of agents and their attitudes are

[6] You needn't agree that Bayesianism cannot do all of these things—arguments about the scope of a model are common and important. But I hope you do accept that these purposes often motivate particular choices of formal framework. I return to competition between models in section 6.

represented as being normatively ideal. In representing real agents as being a certain way, normative models eschew the goal of descriptive accuracy. In this, they are no different from other representations: stylised imagery, caricatures, or heavily touched-up glamour photography. The theory of scientific modelling has the resources to explain this, drawing on the notion of representation-as developed by Goodman (1976) and Elgin (1983, 2010).

The idea is to separate out two parts of our ordinary notion of representation. Consider the example of a famous caricature of Churchill as a bulldog standing on Britain. This is a representation of Churchill, but in a sense it is also a representation of a bulldog (after all, it is a picture of a bulldog with a vaguely Churchillian face). We separate out these two notions by calling the first kind, involving denotation of a target, representation-of; and the second, involving the way in which the target is shown (also known as its secondary subject), representation-as. The formula is: an object X (the drawing) represents a target Y (Churchill) as something, Z (a bulldog).

The Z variable identifies the secondary subject; the kind of representation it is, or what it portrays. I will refer to these genres of representation as Z-representations (e.g., bulldog-representations). As our formula says, Z-representations are objects (like drawings), and what fixes the genre Z is an interpretation. We can think of an interpretation as a function, mapping properties of the object to properties of Z. We associate properties of the caricature (particular lines, shading etc.) with properties of a bulldog (a certain stoutness, folded skin, etc.). Interpretations allow us to talk about Z-representations as "having" Z-properties that, strictly speaking, they do not. (The drawing does not have four legs, it is a drawing.) With something like a caricature, certain properties are highlighted as particularly relevant and the intention is that we impute those properties to the target. Bulldogs are pugnacious, and the caricature highlights this with the stance of the Churchill-dog in the drawing, with the intention that we regard Churchill as pugnacious. This is the final part of representation-as: when there is a target, we can impute highlighted Z-properties to the target.

A number of popular accounts of scientific models agree that they utilise representation-as (Hughes, 1997; Elgin, 2009; Frigg and Nguyen, 2016). We can deploy this same conceptual infrastructure to account for the normative aspects of NFE. Our normative models in NFE are representations of real agents and their beliefs, and they are also agent-representations. These agent-representations are constructed in part to resemble the targets—they have the same kinds of attitudes as real agents, as well as various other features. But they are also idealised in various ways, so that the agents portrayed are dissimilar to real agents. Some of these differences are descriptive idealisations (I will suggest which below). Unlike scientific models, however, some of these differences are normative. The model's results are used to generate claims about the target, but here the claims are normative rather than being descriptive. The real agents are represented as normatively ideal agents.[7]

My account explains why philosophers occasionally say that these are "models of ideal agents"—it is the same usage as calling the Churchill caricature a picture of a dog. Much of our work in formal philosophy involves the manipulation of the model objects (the mathematical structures), in the form of deriving results and interpreting them in terms of the properties of agents. So, far from being a barrier to understanding normative philosophy as modelling, representation is key to understanding how NFE

[7] Note the double-duty that the word "idealised" is doing here. Models are also idealised representations, without the word "ideal" implying anything normative. In the philosophical case, "ideal agents" are normative ideals but we see now that they may also be also descriptively idealised.

generates normative claims that are relevant to real agents.

4.2 Data

Scientific models are constructed using data, typically measurements of the target system such as the records of my observations of the fish pond. The model is constructed to "fit" this data, which is to say that the claims generated by the model about the target should agree, closely but not completely, with the data. (This wiggle room is inserted because data are typically noisy.)

What are the data used to construct a normative model? Titelbaum takes his answer to this to offer the defining feature of normative modelling in philosophy: "I distinguish descriptive from normative models in terms of the data they attempt to fit. Descriptive models attempt to fit descriptive facts about, say, some natural or social phenomenon. Normative models attempt to fit normative facts. Examples of normative facts include prescriptions (you shouldn't act against your own best interests) and evaluations (it's irrational to believe a contradiction), but also general facts involving normative concepts (correct inference requires truth-preservation)." (Titelbaum, b, 2-3).

While I agree with the broad thrust of this, I want to add some additional detail. The "facts" in Titelbaum's parenthetical remarks are certainly what we are attempting to capture, but they are not the inputs to the process of modelling, in the same way that facts about the way the world is are not what is used in the construction of scientific models (Bogen and Woodward, 1988). Rather we fit scientific models to data, which are the result of measurement, or calculation, or what have you. I favour an account on which our philosophical data are judgements—either our considered judgements made directly about normative matters (also called intuitions) or perhaps judgements made on the basis of rigorous argumentation or previous modelling exercises. The process of "considered" judgement is analogous to the preparation and cleaning of empirical data—we are attempting to reduce noise, so that these data give us a clearer picture about the facts. Individual data might be rejected as outliers. All this is analogous to science, but not captured in Titelbaum's fact-based view (though Titelbaum does acknowledge these sources of data.)

Two further differences between Titelbaum's view and my own are relevant. The first is that I insist that models in NFE also rely on descriptive data—capturing trivial features of the agent, such as that they have beliefs, and more relevant information such as the content of those beliefs. The second difference concerns what makes something a normative model. Titelbaum says it is the fact that the model fits normative data. I disagree: any model put to a normative purpose is a normative model.

4.3 Normativity as Purpose

I will now defend this claim that a normative model is any model that is put to normative purposes—evaluation, action-guidance, exploration of putative norms, and perhaps others.

The use a model is put to depends on the purposes of the modeller, so whether a model is normative depends on the modeller. Normative models are thus not of a fundamentally different kind to scientific models, in fact, one can be redeployed as they other. I am predisposed to an account that unifies normative and scientific modelling in this way, for part of the purpose of describing NFE as modelling is to utilise the lessons of philosophy of science to improve our philosophical methods. But I also think that, given the relations between normative and descriptive models that I lay out below, it would be strange and unparsimonious to insist that normative and descriptive models were different enough that they require distinct accounts and distinct methodologies.

To begin, let us note that normativity is not so foreign to science. Physiological models in medicine can be thought of as normative, representing how the body should be, with actual deviations representing illness. Ecological models might represent an undisturbed ecosystem and thereby act as an evaluative standard for assessing the impact of alien species. Architectural models describe how buildings ought to be built. Nonetheless, current work in the philosophy of scientific models does not focus on these normative aspects, preferring representation as a topic of philosophical discussion. So, addressing normative models is of value both to core philosophy of science and to our present metaphilosophical project.

The main reason to regard the normativity of normative models as coming from the purposes of the user is that models can be easily redeployed as normative or descriptive. This happens in engineering, when an architectural model which began as a representation of a to-be-built building becomes a normative model for the construction team: telling them how to build, and providing a standard against which they are evaluated. When this happens, there is no change in the data the model fits, nor is there a change in the representational target of the model (there isn't one). So the model's normativity doesn't come from fitting any normative data, but rather from the user.

We can see this happening in NFE by looking at the decision theory-economics borderland. Following Buchak (2013) we can distinguish four projects of decision theory: construed normatively, decision theory can be used to evaluate or guide actions; construed explanatorily, it can be used to describe or interpret actions.[8] The normative-evaluative use of decision theory involves analysing a decision situation facing an agent, and determining which actions are rational. The normative-action-guiding use involves deploying this process expressly in order to determine which act to undertake. The descriptive-explanatory use and the interpretive-explanatory uses are interested in real, rather than ideal, agents but they differ in their goals. Descriptive theorists are interested in describing observed patterns of behaviour—this is the empirical project of rational choice theory within economics. Interpretive theorists take real agents to be aiming at prescriptions of rationality but failing for various reasons. This theorist seeks to interpret the actions of the agent, as much as possible, as abiding by the rational theory of decision.

The important point is that the very same logico-mathematical apparatus is often deployed in each of these projects (perhaps with different degrees of success). There need be no difference in the mathematical description of a model used in a normative-evaluative mode by philosophers to generate norms of rationality, and in a descriptive-explanatory mode by economists to predict choice behaviour. Economists often say they are using Savage's model, which was motivated normatively. But the economist's project is not normative: they are interested in predicting, explaining, understanding. One might insist, with Titelbaum, that these are normative models because they fit normative data. But I think it more natural to say that they are descriptive models which include some normative data. This inclusion might be justified by regarding a normative property—such as transitivity of preference—as approximately true of real agents who acknowledge the norm.

This movement can happen in either direction: a model can begin life as normative, and later be taken up for descriptive purposes, or vice versa. One easy way to go from

[8] I have relabelled these projects for convenience, taking inspiration from Thoma (2019).

descriptive to normative is to fix a physical system as a reference point for generating norms, and thereby turn a descriptive model of that system—say, Buddha—into a normative model—generating claims of the form "do as Buddha did".[9]

Whatever the origin of the model, what makes it a normative model is that the modeller intends it to be taken as normative.

4.4 Justifying Idealisations Normatively

As I mentioned above, we expect models to fit the relevant data. But "fitting" is a match between outputs and data. How to philosophers get their models to fit some data? One common strategy is to use normative data—judgements about rationality—in their construction. Philosophers in NFE build these judgements into their models in the form of normative properties, like the transitivity of preference and partial belief. I propose that we regard these normative assumptions as a species of idealisation, justified normatively.

They are idealisations in that they distort or leave out properties of the real system, paralleling Galilean or Aristotelian idealisations in science. The idealisations of science, which we can call "descriptive idealisations" to distinguish them, are justified in a variety of ways each related to the purpose of the inquiry. Musgrave (1981) gives a helpful taxonomy of justification. In one common case, a modeller takes an idealisation to have negligible effects: for the purposes of the current investigation it will make negligible difference to distort/exclude this property. 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. 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. Many models contain idealisations justified in different ways.

I suggest that we view presumed normativity as a fourth species of justification. A modeller introduces a distortion, relative to the properties of the real system, because they take it that the system ought to be this way. (Or alternatively they wish to explore the consequences of regarding it as a norm.) Investigating that norm is part of their purpose in building the model.[10] Normativity is a species of justification because the same assumption (e.g., transitivity of preference) can be introduced as a norm (e.g., by a philosophical decision theorist) or as a descriptive idealisation (e.g., by an economist) justified perhaps by arguing that deviations are negligible.

[9] For an example of work doing exactly this—generating norms from a real person—albeit not in a modelling context, see Olberding (2017). ACKNOWLEDGEMENT

[10] Remember that it is the purpose of the modeller which makes the model as a whole a normative model. I am speaking here about particular assumptions or constraints in the model which are included because the modeller takes them to represent norms. A model need not have any of these to count as a normal model—as described with my example of Buddha above.

Consider Earman's model, and the property of monotonicity (P4): degree of belief does not decrease across entailments. An economist might motivate this idealisation by any of Musgrave's three kinds of justification. The philosopher, by contrast, has a fourth option: degree of belief ought not decrease across entailments.[11] When philosophers and economists use "the same model," for normative and descriptive ends respectively, what they are doing is construing these conditions in different ways.[12] There are at least two varieties of this normative justification. Sometimes assumptions are included because the modeller takes them to be norms (they are supported by reliable data, or are firm considered judgements). But at other times, these assumptions are the object of study. The aim is to study how various putative norms hang together, or what a particular norm's consequences are.

I think the models in NFE typically employ both normative idealisations and descriptive idealisations. That is because I don't think that all of the distortions in these models are norms: completeness and continuity of partial belief, or completeness of preference, to name a few. Economic and philosophical models in decision theory rely on many of the same assumptions and there is no need, as I see it, for the philosopher to insist that all of these assumptions are norms. Modelling is a methodology which makes regular use of precisely this sort of descriptive idealisation.

Our normative models may thus contain artefacts, expected side-effects of their use of idealisations. This is one way to understand logical omniscience: a side-effect of certain desirable aspects of this model of partial belief. For example, one might think it is a norm that degree of belief should not decrease over known entailments. But in the standard Bayesian model, the entailment structure is built into the way we represent propositions and so the logic is "objective"—it doesn't depend on the agent's attitudes. Shifting away from using an objective logic is a daunting task—models of bounded rationality are often complex (e.g., Garber, 1984). Instead, a philosopher might use a model in which the agents know all entailments—a descriptive idealisation which, together with the norm, renders them logically omniscient in the sense here discussed. They do this for reasons of simplicity, and so to them this property is not normative and this is marked in the model's key. (Such a model can't separate out the entailments covered by the actual norm about known entailments from those not so covered, and so it not suitable for investigations that are too close to this norm, such as studies of deductive inference.) This behaviour amongst philosophers—of simultaneously disavowing logical omniscience but continuing to use Bayesian tools—is explained by reconstructing their method as modelling. This reconstruction allows us to explain the connection of this kind of work with that of philosophers interested in bounded rationality. The latter group build less idealised models, seeking to remove descriptive idealisations. This increases the complexity of the models, but simplifies their use. The less descriptively idealised the model, the closer we can get to simply "taking its results seriously", for all those results.

[11] Pettigrew calls this norm "No Drop" (Pettigrew, 2016, 2).

[12] However, some economists currently describe their ends as*descriptive* though they justify their idealisations normatively. Their models rely on assumptions like the transitivity of preference, which the economists justify by stating that it is a *norm for preference*. On my account, justifying the idealisation normatively is in tension with using the model for descriptive ends, and represents a conceptual confusion on the economist's part. It would be better to say that the idealisation is approximately true, which in turn might be explained by the fact that transitivity is a norm of preference.

5 Modelling or Bust

Until this point, I have only tried to show that one can regard NFE as modelling. I've focussed on Bayesianism, but the features that I have drawn upon in this characterisation are plainly present in other bits of normative formal epistemology. So why should one use this characterisation? My argument is this: modelling is the rational reconstruction of normative formal epistemology's method on which that method is best justified. NFE does better against its goals if it is modelling. In short, metaphilosophers are faced with a dilemma when characterising the method of NFE: modelling or bust.

I said in the introduction that this is a rational reconstruction and so I will briefly say what I think that is and why such a thing is relevant. The method of rational reconstruction involves presenting some scientific practice, or pattern of reasoning, in a manner which is best suited for justification. It is partly descriptive, in that it is meant to "fit with" the actual practices and conclusions. But it is not an exercise in psychology or sociology. The reconstruction is "rational" in that it contains only the epistemically relevant aspects of the practice (Laudan, 2004, 15). It relies on a kind of principle of charity—the goal is to achieve a reconstruction of the practice on which it does best. For Reichenbach, the aim is that "the rational reconstruction expresses what we mean, properly speaking" (Reichenbach, 1938, 6). Such Reichenbachian reconstructions serve a critical purpose: the aim is to prepare the method under study, presenting it in its best light, before making philosophical recommendations.

I characterise my work as rational reconstruction for two reasons. First, it makes it clear that it is no objection to my project that philosophers don't think that they're modelling, or that they don't consciously employ the concepts and tools I describe. Secondly, I agree with French that when we provide philosophical accounts of scientific theories, models, theorising and modelling we are almost always engaged in a kind of reconstruction (French, 2020, 5). So if we are to discuss what we're doing, methodologically, rational reconstructions are one useful way of conducting that discussion.

5.1 The Landscape of Methodological Possibilities

I claimed that formal epistemologists face a dilemma: modelling or bust. But surely there are many things we could be doing in NFE? In this subsection, I will discuss various ways of characterising the methodology of NFE and show why I view the choice as dichotomous.

First, a comment on the scope of my analysis. Not all of science is modelling, and I don't think all of philosophy is either. Godfrey-Smith (2007) distinguishes the model-based "strategy" of science from an alternative, more direct, method of theorising in biology. Godfrey-Smith's example of direct theorising employs no simplified representations; instead, it examines actual organisms, in their actual circumstances. This work is close to the data, and involves studying real rather than fictional systems. It is synoptic, making progress by systematising knowledge.

So, even if naturalistic philosophy recommends using scientific methods, these needn't be modelling. This direct method, however, is not a good description of NFE. We don't work from close attention to real agents in real situations, and not merely because we have normative aims. Consider *Vices of the Mind* by Quassim Cassam (2019). It is epistemology done "from the ground up"—the theory of epistemic vices is built from a close examination of real cases but the theory is put to normative ends. It is manifestly unlike the formal epistemology discussed here. There is no conflict between the dichotomy

I am working towards for NFE, and the existence of successful non-modelling work outside of NFE.[13]

Modelling is a method rather than a goal. It therefore doesn't conflict with the traditional philosophical project of conceptual analysis, nor with ameliorative analysis, or Carnapian explication. As noted above, models are used to isolate mechanisms or concepts for particular study (Wimsatt, 2007, 15). Models can support any of these projects by providing a setting in which a philosopher demonstrates the value of a particular explication, or a way of isolating the concept being analysed. It is true that models are limited in scope, and so their use might complicate conceptual analysis which seeks to offer a universal analysis. But this is analogous to using models in physics, where the goal is to derive universal laws of nature. All that it requires is care.

Modelling can also support seeking reflective equilibrium. Reflective equilibrium names both a goal and its associated method. Traditionally, reflective equilibrium is attained by testing theoretical principles against our judgements about cases. If there is a conflict between our judgement and the verdict of the principles, we must choose how to make the two consistent. One option is to change the theory, another is to revise or reject a judgement. The aim is to iterate this procedure until we arrive at reflective equilibrium; a point at which our judgements and theory are consistent and remain so in the face of new cases. There is no conflict between this procedure and using a model to generate those verdicts. Care will be needed to ensure those verdicts aren't the result of artefacts, or overly sensitive to arbitrary modelling choices. Models can be a way for philosophers to test the conclusions of various principles without resolving all theoretical questions: by building a model which captures a limited subset of principles in a restricted domain, models can provide a testing ground for particular aspects of the theory.

5.2 The Trouble with Theory

Perhaps the most natural alternative is that the formal structures used by philosophers are intended as theories, rather than models. Though not so common in epistemology, the term "theory" is nearly ubiquitous in ethics and political philosophy, where theories are often explicitly described as being universal and general, consistent, complete, axiomatisable, and decidable (by defenders and detractors of theory—see Roussos (ms)). The rough notion of theory used there is that of a system of claims, centred one or more laws: universal, exceptionless generalisations. On philosophy of science's "syntactic view", theories are precisely such sets of sentences, which are consistent and closed under entailment. This seems to fit the ideals of philosophy: do we not seek universal, necessarily true, laws systemised in such theories? It also fits with practice: revealing inconsistencies and false entailments is a core part of the traditional philosophical practice of putting forward a conjecture only to have it refuted by a colleague's counter-example. (For the sake of precision: from here on I use "theory" in this way and "theorising" to refer to the practice of theory-building.)

Applied to NFE, this generates a rational reconstruction quite different from mine. If we are theorising, then the formalism of NFE is merely a language used to present descriptions and claims which are meant to apply universally (Titelbaum (b) calls this the "abbreviated description" view of formal philosophy). Work in NFE is intended to

[13] Similarly, if there is a literature I am not aware of, in which formal epistemology for normative ends is done "ground up" from detailed observations of real agents, then it is simply not part of my target here.

be of universal scope, and its "laws" should be read literally and assessed for truth.

Offered this alternative, many epistemologists might agree with Rosanna Keefe (2000), who discusses modelling disparagingly in her book on vagueness. Rather than presenting an idealised, partial representation she aims at a true description of the phenomenon of vague language. Accordingly, she says, her methodology is not that of the modeller and she is explicit that every implication of her work is to be taken literally. As Keefe notes, it is not open to her to tell us to disregard certain parts of her mathematical framework as artefacts, or to isolate her account of vagueness from other accounts of linguistic functioning. Not for her the "Get Out of Jail Free" card of modelling. The success conditions for her work are straightforward: truth; her work is open to refutation by counterexample, by design (Keefe, 2000, Ch. 2).

But adopting a theory view of NFE would put enormous pressure on the many idealisations it employs. The many criticisms of Bayesianism threaten to undermine the whole project, if we view it as a theory in the business of generating all and only truths. To illustrate these problems, I will walk through the following decision tree. First we decide: modelling or theorising. Now consider all the assumptions I labelled "idealisations" above. What are these: norms, or not norms? An answer must be provided for each. If any are not norms then the theorising view is, I think, doomed. If all are norms, the project is beset by counterexamples, in the form of considered judgements that agents who violate these properties violate no norm. Thus, working back up the tree, the choice is: modelling or bust.

So, let start with the first node, and suppose that we are theorising. Bayesianism describes agents as logically omniscient, with complete and precisely commensurable partial beliefs. Bayesian agents therefore have priors for every proposition—a feature which is a necessity when we introduce the Bayesian theory of learning by conditioning. For any possibility, they have a prior attitude to it encoding exactly how surprised they are that it came about. Bayesian agents have no computational limitations, no limit to the number of beliefs they hold. These are all false descriptions of real agents. But on the theorising view, each must be taken seriously, which means each must be interpreted as a normative statement about the real agents that Bayesianism is intended to be relevant to. There is perhaps some wiggle room in the interpretation—one might insist that conditional priors represent evidential standards rather than beliefs (Meacham, 2016; Titelbaum, a)—but once the interpretation is fixed, every consequence must be examined.

These criticisms are well known (for classic discussions, see Glymour (1980) and Earman (1992)), but let us review some issues with describing these all as norms, regarding logical omniscience. What prevents us from meeting the demands of Earman's (A2), (P2), and (P4) is our cognitive limitations. We find it difficult to determine whether sentences express tautologies, or whether two sentences express the same proposition. Some of these tasks are very difficult indeed—capturing the attention of the most talented mathematicians or requiring thousands of hours of computational time on algorithms built by our best computer scientists. These limitations are not usually thought of as subject to the norms of rationality—it strikes us as plainly wrong to say that mathematicians were simply irrational for not knowing whether Fermat's Last Theorem was true. Rationality is in the business of telling us how our beliefs should fit together, or change over time, or how we should act given our beliefs. It isn't in the business of telling us that we ought to have vastly different capacities, nor that we ought to have specific beliefs—such as full credence in logical truths.

Every problem for Bayesianism must be examined in this way. Any non-normative and false elements are intolerable on the theory view. Now one might be tempted to say that Bayesianism is a work in progress, that a future Bayesianism will contain only normative or true elements. But this would be to give up on a lot: Bayesianism would be relegated to the status of a promising research programme, rather than being one of our dominant theories of rationality. It would be unclear why anyone should take seriously the current results, enmeshed as they are in logical omniscience, completeness, continuity, full awareness, and so on.

In short, if Bayesianism is a theory, it is a bad one.

Two objections are worth considering here. I have been treating Bayesianism as a theory of very wide scope, as a "guide to life." But could it not be more local, or restricted in some way? Perhaps the "global" Bayesianism is doomed, but more local theories can succeed. This is one of a number of "escape strategies" which point away from theorising and towards modelling. For example, one might say that Bayesianism is only meant to be applied to situations in which agents do have complete and commensurable beliefs. This is a domain restriction, and characterises completeness and commensurability as domain idealisations. This is how the decision theorist Itzhak Gilboa (2009, 51-2) justifies the completeness of preference. The theorist can proceed in this way for every problem with Bayesianism, but the result is a "theory" about agents who look nothing like us, and which is applicable only in unusual situations. This is not much like the picture of universal and exceptionless generalisation we started out with.

Or perhaps, as a reviewer suggested, theorists are merely employing some division of labour. I take it that the idea is to divvy up the task of relaxing these idealisations. Theorists don't endorse logical omniscience, it merely isn't what they're working on. This seems, to me, straightforwardly a modelling strategy. Idealisations are not "endorsed" in science either and, as Musgrave noted, scientists often adopt idealisations only heuristically. But our results are not easily "detachable" from our idealisations, the justifications for employing them, and the restrictions they impose. Theorising-plus-division of labour does not seem to provide any method for managing these useful falsehoods or their effects. We cannot take literally all the results of the theorist who ignores logical omniscience. So how should we treat them? The answer, surely, is that given in this paper: the answer of the modeller. (Managing idealisations will be discussed further in section 6.)

The choice facing epistemologists is this. You can insist that you are theorising, that your results are intended to be read literally, and that the standard for their evaluation is truth. In this case, either every assumption you employ is normative or most efforts in NFE contain straightforward falsehoods and offer no method for sifting the true from the false. Alternatively, you can say that you are modelling.

6 Costs and Consequences

Accepting that NFE is modelling is not without costs. In this section, I examine some of what we will have to give up and what we will have to change, once we accept that we are modelling.

6.1 Securing Normative Inferences

I will start with the question raised at the end of the previous section. How can we know that the normative inferences of NFE are "secure" in the face of their dependence on non-normative idealisations?

The problem concerning us here is that some of our normative conclusions depend necessarily on these assumptions. Supposing that completeness is neither a normative standard nor an approximately true description of partial belief, what does that mean for Probabilism? More generally, do idealisations pose a threat to our normative inferences?

Important strategies for determining the impact of idealisation on scientific models go under the headings "sensitivity analysis" and "robustness analysis". The former typically refers to varying some numerical parameters in a model and studying the changes in the results, while the latter typically refers to comparing the results of differently idealised models. One particular worry for idealising assumptions is the possibility of "knife edge" results—highly unstable results generated by the nature of the idealisation assumptions (e.g., when they introduce a symmetry). Here, small variations lead to very large changes of output, undermining the value of the model as a tool for studying a real system, which is non-ideal. I will examine how we might deploy these strategies in philosophy.

The first strategy for securing our normative conclusions is de-idealisation. Suppose we are interested in some conclusion from a highly descriptively idealised model. We construct a new model with the same normative assumptions but fewer/less extreme descriptive idealisations, and see whether the conclusion of interest emerges. Such a model will typically be harder to build, and harder to use—idealisations, remember, simplify analysis. But suppose that it is done, and that the conclusion re-emerges: this lends some support to the claim that it has normative force. We could build up this support inductively by considering successively less idealised models. This is a relatively simple sounding strategy, but I am pessimistic about its success: in science de-idealisation is often very difficult or impossible.[14]

More often we will have to put up with our idealisations. Scientific models work well, often because of idealisations, but their use involves an acknowledgement of their limitations and fallibility. Their value is in allowing the modeller to capture the main features of a system's behaviour, driven by the most important underlying factors. In complex systems this is remarkable, but it comes at a cost. Sometimes that cost is precision: the model captures a general pattern, but cannot make precise predictions. Sometimes, it is generality: the model works well in this particular case, but not outside of it. Scientists determine these restrictions by considering the nature of their idealisations. Philosophers, too, ought to do so. This may result in regarding results as normative under certain conditions, or within a certain domain. The philosopher might conjecture that the norm holds more widely, but no more. This is in contrast to current practice in which a particular mathematical argument might be presented as establishing the norm of Probabilism.

One option is to study the role of the descriptive idealisation mathematically, and thus bolster the argument that the result is normative outside of the domain of the model. In simple cases, this looks like varying the value of a numerical parameter: if the idealisation is no friction, represented by $k = 0$, we might study the behaviour of the model for small values of k. If our descriptive idealisation takes this form, this may be a useful strategy. I expect this to be more difficult in philosophy, where our assumptions are rarely framed as numerical parameters. How exactly to perform this sensitivity analysis will depend on the case, but I will offer two examples to illustrate some basic concepts.

First consider the famous paper on the interpersonal comparability of individual wel-

[14] This process of de-idealisation was proposed by McMullin (1985) as a way of neutralising the challenge idealisations pose to realism in the philosophy of science. However, Morrison (2005) has argued convincingly that this de-idealisation is often impossible in modern physics.

fare by Sen (1970). He notes that ideal theory results assume comparability, which is philosophically suspect, but that the opposing literature makes the equally extreme assumption that individuals are not comparable at all. Sen instead studies different points on the spectrum between full interpersonal comparability of welfare and incomparability, in order to determine which desirable results continue to hold at each point.[15] The genius of this paper is in taking seriously the observation that we happily make some welfare comparisons, by finding a way to construct a spectrum of partial comparability rather than assuming full comparability.

Second, recall that Arrow's impossibility theorem states that, when voters have three or more options, there is no way to convert the preferences of individuals into a collective preference ranking while also meeting a set of criteria that are taken to represent normative ideals for democracy (Arrow, 1951). Of particular interest in the subsequent literature was the risk that the social ordering would contain cycles. In a well-known paper Gehrlein (1983) demonstrated via simulations that the larger the population, the greater the chance of cycles. This caused much concern that democracy might be irrational. In tension with this were empirical results showing that cycles are in fact relatively rare (Mackie, 2003). The explanation of this tension is that the formal result was an artefact. Gehrlein made use of an "impartial culture" assumption, itself the conjunction of a normative idealisation (universal domain - a norm of democracy that we set no restrictions on which personal preferences are allowed) and a descriptive idealisation (equiprobability - that each of these preferences is equally likely in a population). Gehrlein's increasing likelihood of cycles is a knife edge result: sensitive to the precise equiprobability of preferences. Christian List proved that "given suitable systematic, however slight, deviations from an impartial culture, the probability that there will be a cycle under pairwise majority voting vanishes as the number of individuals increases" (List and Goodin, 2001, A4). He concludes: "An impartial culture is a rather unstable limiting case."

Careful analysis of the relation between idealisations is another fruitful strategy. The "coherent extendibility" thesis provides a good example of this. The version I will consider here concerns partial belief, and the interplay between transitivity and completeness. Suppose incompleteness is rational, so that an agent might fail to compare propositions A and C in terms of comparative confidence. But suppose that they judge that A is more likely than B, and B more likely than C. Transitivity demands that they judge A more likely than C. But, since we are permissive about incompleteness, we don't want this judgement. What this calls for is a revision of the norm of transitivity, and a reinterpretation of the role of completeness. Failures of transitivity due to incompleteness are no problem, we now say. But, should the agent's gaps be filled in, the extension of their attitude to a complete attitude ought to be coherent (see (Jeffrey, 1992, 85), (Joyce, 1998, 103)). An agent whose current, incomplete, belief state commits them to violating transitivity if they complete their attitude is at fault, rationally speaking.

The preceding paragraph is a way of understanding the idealisations of transitivity and completeness which feature in a standard Bayesian model. It is a "key", in the sense of Frigg and Nguyen: it tells us to read transitivity in this qualified way, and to regard completeness as neither simply a descriptive idealisation nor as a norm itself. Rather, it facilitates a focus on other idealisations: the normative ones. The model represents not the agent's actual beliefs, nor the normative requirements on those current beliefs, but the normative requirements on a completion of those beliefs. So it can't be read as generating norms that are straightforwardly applicable to real agents. Special care must

[15] ACKNOWLEDGEMENT

be taken with any differences between the model and reality that result directly from this completion. Some aspects of the credal representation may be non-unique given the coherent extendibility interpretation; for example, there may be different completions compatible with the agent's real comparative confidences, each of which would generate different numerical credences. If so, we should not rely too closely on the precise values of these credences. However, we can straightforwardly rely on aspects of the representation which come directly from the agent's actual comparative confidence, such as inequalities between credences for propositions towards which they have an attitude. Due to these limitations, if one is interested in situations where the incompleteness is thought to be important, this model may be of limited use.

The emerging view is a cautious one. Certainly more cautious than current practice, and significantly more qualified than one recent discussion of the success of normative models. In a recent paper, Beck and Jahn (2021) propose normative models secure their normative inferences simply by fitting the data, a set of independently justified normative judgements. A model which does so identifies a pattern to these judgements, which it then projects onto new cases. The role of normative models is thus to extend normative justification from familiar to novel cases. For Beck and Jahn, there need not be any normative assumptions involved in building a model: a black-box machine learning model could be a successful normative model, so long as the data it is fitting is a set of normative judgements. The modeller's purposes play no role in their account and normative idealisations, as I define them, are discussed only in a competing account that Beck and Jahn ultimately reject.

While their account has some appealing features I think it is unsatisfactory for three reasons. Beck and Jahn seem to take evidence that a model works in known cases—i.e., that it fits data—as reason to believe it will work in new circumstances. The first issue is that they don't offer a justification for this move. In science such inferences are justified by a kind of inference to the best explanation about the model capturing a mechanism or causal capacity which generated the pattern and which scientists expect to be stable. No parallel argument is offered by Beck and Jahn (for discussion of a this kind of stability and projection in normative modelling, see Roussos (ms)). Second, Beck and Jahn give us no way to identify a domain in which a model will produce reliable inferences. On my account, idealisations and purposes set the scope and, as I will discuss below, descriptive idealisations limit a normative model's scope. Third, while their account might tell us when an inference is secure, but it doesn't allow us to generate explanations. On my account, normative idealisations are the normative inputs, which then naturally play a role in explaining the derivative norms that emerge from the model. As Beck and Jahn neglect purpose and idealisation, the resources I use for these purposes, I don't see how they could account for them.

6.2 Purpose, Scope, and Criticism

Scientific models are purpose-specific, with restricted domains of applicability. Does this also apply if we choose to think of NFE as modelling? In a weak sense of purpose-specificity, this might seem trivial. We are discussing agential models of rationality, built to explore the rationality conditions on partial belief. That fills the basics of "purpose"—it tells us what the model is a model of, and what it is trying to do. But does this purpose also lead to modelling choices that restrict the model's usefulness in answering some questions, giving it limited scope? Are our models evaluated not on their truthfulness (the ultimate aim of our inquiry), but instead on their adequacy for purpose (the immediate aim of model-based inquiry)? I argue yes. I will demonstrate two further ways in which NFE models are purpose-specific. The first related to high-level purposes, and the second relates to idealisations.

To begin, let's consider the high-level purposes of epistemologists engaged in NFE. I will name a few distinct projects that philosophers use models for, and doubtless there are others. We might have some norms, like transitivity and monotonicity, and want to see what they imply for the structure of partial belief. Or, the norms might be under investigation: we test a candidate norm by seeing whether it generates counter-intuitive consequences. We might test a set of norms, to see if they are consistent (a common practice in social choice theory). Each of these aims will imply virtues that models should have to facilitate them.

Consider again the idealisation of logical omniscience. Regardless of whether it is a norm, the Bayesian models we have been discussing have a natural domain restriction: they are no good for modelling cognitively limited agents reasoning in the face of logical or mathematical uncertainty. I think it is open to Bayesians to say: this is a model of empirical uncertainty, it is not intended to be used for other purposes. It comes with a key which tells users to disregard claims which are about logical uncertainty. The boundaries are going to be fuzzy, however, and so we should expect disagreements about what counts as the legitimate domain of Bayesianism qua model of empirical uncertainty. So, evaluating such models requires paying attention to these purposes. This gives us a way of focusing on which truths the model is aiming at, and what degree of closeness to the truth is required for those domains.

As you can now see, 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. Counterexamples count against a model directly only when they are within its scope. It it therefore of first importance to delineate these scopes, explicitly and upfront, when we model. Counterexamples from outside a model's scope can count against it indirectly, when models are being compared. If a second model performs better on the shared purpose and has wider scope, then it will be favoured. But this is quite rare in science—models often cannot do exactly the same things as one another. So we may have to get used to a landscape of different models, each built for different purposes. Our understanding of the epistemic domain is enriched by having multiple perspectives, even if they are limited and contrary, and even if none is an obvious candidate for the one true theory of the epistemic.

This will be disappointing to many who seek a unified framework which delivers universal laws. But to them I say: there is a rich literature in the philosophy of science about the relation between models and theory, the development of laws, and strategies for using multiple models to this end. This is an area ripe for further metaphilosophical work. But for the time being, it clearly calls for changes to how we practice our philosophy. Let me end by summarising three common modes of arguing which need to be treated with care if we regard NFE as modelling.

- Property X appears in our best account of rational partial belief. Therefore, agents are rationally required to have property X. (The "argument for probabilism from representation theorems" employs this move—e.g., in Maher (1993), and see (Hájek, 2008; Konek, 2019) for discussion.)

One we replace the term "account" with "model" it becomes clear we need to be careful. In the scientific case, realist inferences from discovered properties of the model to the target must be motivated with reference to their (in)dependence on idealising assumptions. Similarly, in the normative case, not all properties in the model are going to count as normative. Ignorance of a model's key makes it very difficult to cogently criticise. This leads to a methodological norm for modellers: be clear about what you regard as an artefact, and what you intend to be imputed to the target.

- Property X appears in your account. Property X is absurd, so your account is false. (Glymour (1980) repeatedly deploys this move.)

  As above, we now see that useful models may contain plainly false idealisations and generate odd artefacts, which must not be used when generating claims about the target. But modellers who analyse the interplay between idealisations and work with clear keys have little to worry about.

- Your account doesn't work in case Y. Y is a counterexample, so your account is false. (Very common.)

  Models have a domain of applicability, so each "counterexample" must be checked against this domain. Objections irrelevant to the model's intended purpose have no bite. Instead, they motivate for a different model to be developed (perhaps to handle just those cases, or to expand the scope). Working out the boundaries of applicability for different philosophical models is a research area deserving of more attention.

I presented a dilemma for epistemologists engaging in formal normative work: modelling or bust. The "bust" horn of the dilemma itself offers a two possibilities, both of which I regard as unattractive. Either NFE must rely exclusively on normative assumptions when building its formal representations, or it must accept that its numerous falsehoods undermine the project relative to its own goals. Better to regard NFE as modelling. Once we do, we can identify room for improvement in our methodology—starting with the discussion in this section. Finally, accepting that we are modelling in NFE need not mean giving up on the final goal of true exceptionless generalisations. I mean to remain neutral on whether our final endpoint will be a true and unified theory of the epistemic, or a collection of domain-specific models. But whatever the goal and whatever its chances of success, true exceptionless generalisations are unlikely to be generated by any single modelling study. For those who seek theory, more methodological work is required to develop tools to take us from our models towards that more unified theory of the epistemic.

Acknowledgements

This article grew out of a conversation with Anna Mahtani about what we're up to when we do formal epistemology; my thanks for that inspiration and many helpful conversations. It improved significantly during review, so sincere thanks to the reviewers for BJPS. Thanks also to the "Working Models" group: Margherita Harris, Sam Rijken, Lorenzo Sartori, and Philippe Verrault-Julien. Mike Titelbaum and Liam Bright gave helpful feedback at the start and end of this project respectively. Thanks to audiences at the LSE Choice group, LSE FERG, INEM Como 2020 Summer School, and FEW 2020.

Special thanks to James Nuyen who read about four completely different versions of this over four years. My research was supported in part by the Riksbankens Jubileumsfond (M17-0372:1 Climate ethics and future generations).

Joe Roussos
Institute for Futures Studies
Holländargatan 13
Stockholm, Sweden
joe.roussos@iffs.se

References

Kenneth J. Arrow. *Social Choice and Individual Values.* Yale University Press, 1951.

Lukas Beck and Marcel Jahn. Normative Models and Their Success. *Philosophy of the Social Sciences*, 51(2):123–150, 2021. Publisher: SAGE Publications Inc.

James Bogen and James Woodward. Saving the Phenomena. *The Philosophical Review*, 97(3):303–352, 1988. Publisher: [Duke University Press, Philosophical Review].

Lara Buchak. *Risk and Rationality.* Oxford University Press, 2013.

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

Nancy Cartwright. *Nature's Capacities and Their Measurement.* Oxford University Press, 1989.

Quassim Cassam. *Vices of the Mind, From the Intellectual to the Political.* Oxford University Press, 2019.

Mark Colyvan. Idealisations in normative models. *Synthese*, 190(8), 2013. URL `http://www.jstor.org/stable/41931906`.

Henk de Regt. *Understanding Scientific Understanding.* Oxford University Press, New York, 2017.

John Earman. *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory.* MIT Press, Cambridge, MA, 1992.

Catherine Z. Elgin. *With Reference to Reference.* Hackett, Indianapolis and Cambridge, 1983.

Catherine Z. Elgin. Exemplification, Idealization, and Understanding. In Mauricio Suarez, editor, *Fictions in Science: Essays on Idealization and Modeling*, pages 77–90. Routledge, London, 2009.

Catherine Z. Elgin. Telling Instances. In Roman Frigg and Matthew C. Hunter, editors, *Beyond Mimesis and Nominalism: Representation in Art and Science*, pages 1–18. Springer, New York, 2010.

Steven French. *There Are No Such Things As Theories.* Oxford University Press, Oxford, 2020. Publication Title: There Are No Such Things As Theories.

Roman Frigg and Stephan Hartmann. Models in Science. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

Roman Frigg and James Nguyen. The fiction view of models reloaded. *Monist*, 99(3): 225–242, 2016.

Roman Frigg and James Nguyen. *Modelling Nature*. Number 427 in Synthese Library. Springer Nature, 2020.

Roman Frigg and James Nguyen. Mirrors without warnings. *Synthese*, 198(3):2427–2447, 2021.

Daniel Garber. Old Evidence and Logical Omniscience in Bayesian Confirmation Theory. In John Earman, editor, *Testing Scientific Theories*. University of Minnesota Press, 1984.

William V. Gehrlein. Condorcet's paradox. *Theory and Decision*, 15:161–197, 1983.

Konstantin Genin and Franz Huber. Formal Representations of Belief. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2021 edition, 2021.

Ronald N Giere. How models are used to represent reality. *Philosophy of Science*, 71: 742–52, 2004.

Itzhak Gilboa. *Theory of Decision under Uncertainty*. Cambridge University Press, Cambridge, 2009.

Clark N. Glymour. *Theory and evidence*. Princeton University Press, Princeton, 1980.

Peter Godfrey-Smith. Theories and Models in Metaphysics. *The Harvard Review of Philosophy*, 14(1):4–19, 2006.

Peter Godfrey-Smith. The strategy of model-based science. *Biology & Philosophy*, 21(5): 725–740, February 2007.

Peter Godfrey-Smith. Metaphysics and the philosophical imagination. *Philosophical Studies*, 160(1):97–113, 2012.

Nelson Goodman. *Languages of Art*. Hackett, Indianapolis and Cambridge, 1976.

Alan Hájek. Arguments For, or Against, Probabilism? *The British Journal for the Philosophy of Science*, 59(4):793–819, 2008.

R. I. G. Hughes. Models and Representation. *Philosophy of Science*, 64:S325–S336, 1997.

Richard Jeffrey. *Probability and the art of judgment*. Cambridge University Press, Cambridge, 1992.

James M. Joyce. A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*, 65: 575–603, 1998.

Rosanna Keefe. *Theories of Vagueness.* Cambridge University Press, Cambridge, New York, 2000.

Jason Konek. Comparative Probabilities. In Richard Pettigrew and Jonathan Weisberg, editors, *The Open Handbook of Formal Epistemology*, pages 267–348. PhilPapers Foundation, 2019.

Larry Laudan. The Epistemic, the Cognitive, and the Social. In Peter Machamer and Gereon Wolters, editors, *Science, values, and objectivity*, pages 14–23. University of Pittsburgh Press, Pittsburgh, 2004.

Hannes Leitgeb. Scientific philosophy, mathematical philosophy, and all that. *Metaphilosophy*, 44(3):267–75, 2013.

E Lemmon and G Henderson. Is there only one correct system of model logic? *Proceedings of the Aristotelian Society*, 33:23–56, 1959.

Christian List and Robert E. Goodin. Epistemic Democracy: Generalizing the Condorcet Jury Theorem. *Journal of Political Philosophy*, 9(3):277–306, September 2001.

Gerry Mackie. *Democracy Defended.* Cambridge University Press, 2003.

Patrick T. Maher. *Betting on Theories.* Cambridge University Press, 1993.

Ernan McMullin. Galilean idealization. *Studies in History and Philosophy of Science Part A*, 16(3):247–273, September 1985. Publisher: Pergamon.

Christopher J. G. Meacham. Ur-Priors, Conditionalization, and Ur-Prior Conditionalization. *Ergo, an Open Access Journal of Philosophy*, 3, 2016.

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

Mary S. Morgan and Margaret Morrison, editors. *Models as Mediators.* Cambridge University Press, Cambridge, 1999.

Sydney Morgenbesser, editor. *Philosophy of science today.* Basic Books, New York, 1967.

Margaret Morrison. Approximating the real: the role of idealizations in physical theory. In Martin R Jones and Nancy Cartwright, editors, *Idealisation XII: Correcting the Model*, volume 86, pages 145–172. Amsterdam, Rodopi, 2005.

Alan Musgrave. 'Unreal Assumptions' in Economic Theory: The F-Twist Untwisted. *Kyklos*, 34(3):377–87, 1981.

Amy Olberding. *Moral Exemplars in the Analects: The Good Person is That.* Routledge, 2017.

Wendy S. Parker. Confirmation and Adequacy-for-Purpose in Climate Modelling. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 8, 2009.

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

Richard Pettigrew. *Accuracy and the Laws of Credence.* Oxford University Press, Oxford, 2016.

Hans Reichenbach. *Experience and Prediction.* University of Chicago Press, 1938.

Joe Roussos. Modelling in normative ethics. *manuscript*, draft date: May 2021.

Amartya Sen. Interpersonal Aggregation and Partial Comparability. *Econometrica*, 38 (3):393–409, 1970. Publisher: [Wiley, Econometric Society].

Paul Teller. Twilight of the Perfect Model Model. *Erkenntnis*, 55(3):393–415, 2001.

Johanna Thoma. Decision Theory. In Richard Pettigrew and Jonathan Weisberg, editors, *The Open Handbook of Formal Epistemology.* PhilPapers Foundation, 2019.

M. G. Titelbaum. *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief.* Oxford University Press, Oxford, 2012.

Michael G. Titelbaum. *Fundamentals of Bayesian Epistemology.* Oxford University Press, a.

Michael G. Titelbaum. Normative modelling. In J. Horvath, editor, *Methods in Analytic Philosophy: A Contemporary Reader.* The PhilPapers Foundation, b.

Michael Weisberg. Three Kinds of Idealization. *The Journal of Philosophy*, 104(12): 639–659, 2007a.

Michael Weisberg. Who Is a Modeler? *The British Journal for the Philosophy of Science*, 58(2), 2007b.

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

Timothy Williamson. Must Do Better. In Patrick Greenough and Michael P. Lynch, editors, *Truth and realism*, pages 177–188. Clarendon Press ; Oxford University Press, Oxford : New York, 2006. OCLC: ocm65201292.

Timothy Williamson. Model-building in philosophy. In Russell Blackford and Damien Broderick, editors, *Philosophy's Future: The Problem of Philosophical Progress.* Wiley, Oxford, 2017.

William C. Wimsatt. *Re-engineering philosophy for limited beings.* Harvard University Press, Cambridge, MA, 2007.

Audrey Yap. Idealization, epistemic logic, and epistemology. *Synthese*, 191(14):3351–3366, 2014. ISSN 0039-7857. URL `http://www.jstor.org/stable/24026192`. Publisher: Springer.