Normative Modeling

Michael G. Titelbaum
University of Wisconsin–Madison

Abstract: By now we are familiar with scientific models of descriptive domains. But might we also model clusters of normative truths? In this piece I first identify elements central to all modeling efforts: modeling frameworks, interpretations, and domains of applicability. Then I consider some advantages and disadvantages of normative modeling.

Keywords: models, modeling, methodology, normative, AGM, Bayesianism

1. Introduction

An article in the sciences will sometimes shift from the description of natural phenomena in words to the presentation of symbols and equations. A reader encountering such a shift for the first time might assume that the author has simply continued the description, but by different means. Perhaps certain locutions become repetitive, so the author has introduced an acronym or symbol as an abbreviation. Describing complex numerical relationships in words can grow cumbersome (as in the work of such early algebra practitioners as Al-Kwarizmi). So the equations may simply be further direct descriptions of the phenomena, presented more efficiently. Call this the Abbreviated Description interpretation of equations and symbols in scientific writing.

Abbreviated Description is undoubtedly the correct interpretation of *some* equation and symbol use in the natural and social sciences. But philosophers of science have suggested an additional interpretation: Instead of directly describing phenomena in the world, equations and symbols sometimes describe formal models. A formal model is an abstract structure, typically composed of mathematical and/or logical entities such as numbers and sets. The scientist describes a model, then links that model to worldly phenomena through various principles.

An article in decision theory, Bayesian epistemology, logic, or the philosophy of language will sometimes shift from the description of principles or norms in words to the presentation of symbols and equations. One might apply the Abbreviated Description interpretation in these cases as well. But is there room for an additional interpretation? Might a modeling methodology be employed in normative disciplines, just as in the sciences?

This essay explores the possibility of normative modeling. I will consider what's involved in normative modeling, and what some of the advantages and disadvantages of this methodology might be.[1] While formal tools have often been applied to normative pursuits, and while many of these applications could be interpreted as instances of modeling, normative modeling has rarely been explicitly discussed.[2] So this piece will be largely exploratory. I hope that it will spark further discussion of normative modeling and a comparison of that methodology to other methodologies formal laborers might employ.

2. Descriptive vs. normative modeling

For my purposes, the distinctive feature of a modeling methodology is that instead of directly describing, predicting, assessing, etc. the target of study, the modeler constructs a separate object (a model), which then does this work by being placed into particular relations with that target. (Think of the difference between describing a mountain range in words and providing a topographic map.) In scientific practice there are many kinds of models; computer algorithms, mathematical structures, and even concrete objects can serve as models.[3] I will focus exclusively on formal models, abstract structures composed of mathematical/logical entities such as numbers, sets, vectors, etc. (The kinds of structure that can act as a "model" in the logician's sense.[4])

A model is designed to fit some body of data. Roughly speaking, the modeler starts with her data, and attempts to fit a model to it. Once she finds a model that fits the data, she can use that model for further purposes, such as explaining features of the data or predicting further data that have not yet been secured. 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.[5] Examples of

[1] The ideas in this essay develop material presented in my (2013). I would also like to acknowledge the influence of Weisberg's (2013) on my general understanding of modeling, though he works strictly with descriptive examples.

[2] Exceptions include Colyvan (2013) and Williamson (2017). While Colyvan focuses on different questions than I will, his overall approach is not too distant from mine; I will mention connections as we go along. Williamson, on the other hand, understands models fairly differently than I do. Williamson has what Weisberg would term a "fictions account" of modeling, on which a model is an invented entity of the same type as the system being modeled (compare Godfrey-Smith 2006). Williamson writes that "a model of something is a *hypothetical example* of it." (emphasis in original) As we'll see, the formal models I'm interested in are abstract structures that need not have any interesting ontological commonalities with what's being modeled.

Williamson also focuses on *philosophical* models, while I will focus on *normative* models in general. Not all philosophical models are normative (Williamson discusses examples from metaphysics), and not all normative modeling takes place in philosophy (I'll mention cases from economics and linguistics).

[3] Weisberg presents an excellent general typology.

[4] Suppes (1960) famously attempted to assimilate formal modeling in the sciences to something like modeling in the logician's sense. I won't evaluate that strategy here.

[5] Compare Colyvan: "Normative models are not supposed to model actual behaviour or explain actual behaviour; rather, they are supposed to model how agents *ought* to act." (2013, p. 1338, emphasis in original) While I've put the point in terms of normative *facts*, expressivists may deny that there are

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).[6]

It's usually clear-cut whether a particular piece of formal machinery is meant to engage descriptive or normative phenomena. But there are ambiguous cases. Formal logic, of the sort we teach in introductory logic courses, is sometimes presented as a science of correct deductive inference, in which case it's normative.[7] Other times logic is couched as the investigation of such concepts as consequence and consistency. If these are read as having no intrinsic normative content,[8] the pursuit is descriptive.

Linguistics and the philosophy of language are often cast descriptively, though they can be read normatively. Linguists and philosophers of language often present data about sentences that are acceptable to native speakers and sentences that are "marked", then try to fit a formal model to those data. Chomskyans read this project as tracking the workings of a cognitive language module, and attempting to predict that module's future deliverances. This is a descriptive undertaking.[9] (The competence/performance distinction then explains why actual speech may fail to satisfy a "correct" linguistic model.) On the other hand, when Stalnaker in "Assertion" presents "some principles that are useful for explaining regularities of linguistic usage", he writes that

> They are not intended as empirical generalizations about how particular languages or idiosyncratic social practices work. Rather, they are proposed as principles that can be defended as essential conditions of rational communication, as principles to which any rational agent would conform if he were engaged in a practice that fits the kind of very abstract and schematic sketch of communication that I have given. (1978/2002, p. 154)

Stalnaker clearly intends to capture normative principles here, and we may assume that the formal structures he developed in this and other work had a similar intent. Similarly, after Grice states a number of his famous maxims in "Logic and Conversation", and casts them in terms of what he "expects" of interlocutors, he writes, "I would like to be able to think of the standard type of conversational practice not merely as something that all or most do *in fact* follow but as something that it is *reasonable* for us to follow, that we *should not* abandon." (1975, p. 48) If projects in linguistics and the philosophy of language are read normatively, then their formalisms may be instances of normative modeling.

such facts. In that case normative models will aim to fit whatever is expressed by sentences like "Killing is wrong."

[6] At this point one might begin to worry how we acquire such normative data—what is the epistemology of these normative facts? I will return to this concern later.

[7] Harman (1986) criticizes this position.

[8] Though the very notion of logical consequence may be seen as normatively loaded. In their exploration of the concept of logical consequence, Beall and Restall write, "Logical consequence is *normative*. In an important sense, if an argument is valid, then you somehow go *wrong* if you accept the premises but reject the conclusion." (2006, p. 16, emphases in original)

[9] Corpus studies (from sources like the internet) are another sign that a philosopher of language is proceeding descriptively.

Let us return, however, to clear-cut cases. The natural and social sciences are clearly descriptive. Decision theory, game theory, and various formal epistemologies (Bayesian epistemology, belief revision, ranking theory) are clearly normative.[10] Here it's a bit distracting that the latter theories are sometimes presented as describing the beliefs or decisions of "ideal agents". But discussion of ideal agents is clouded by an ambiguity among different types of idealization.[11] One type of idealization simplifies for the sake of tractable calculation or analysis—as when physicists work with an idealized frictionless plane. An ideal agent, on the other hand, is ideal in the sense of doing what nonideal agents *ought* to do, or of being better than nonideal agents along some evaluative dimension. (There's no sense in which a frictionless plane is *better* than a real plane.) Thus formal models of ideal agents are normative models as well.[12]

3. Elements of modeling

What's required to apply a modeling methodology? First, we need a modeling framework. A modeling framework is a template, or blueprint, for building individual models of particular situations. To see what I mean, let's start with an example of a descriptive modeling framework. (Later I'll introduce a couple of examples of normative modeling frameworks, so we can compare those to this descriptive example.) I think of the "Lotka-Volterra model" of predation (Lotka 1956, Volterra 1926) as a modeling framework specified by the following coupled differential equations:

$$dV/dt=rV-(aV)P$$
$$dP/dt=b(aV)P-mP$$

These equations contain four parameters: r, m, a, and b. If we set those parameters to particular real-number values, we obtain a model. Usually a specific model is constructed to represent some particular real-world situation. In setting the parameter values of the model, we are guided by an interpretation of the framework that indicates (among other things) which quantities from the world to use as parameter settings in the model.

Thus to engage in modeling we need a modeling framework and an interpretation of that framework, which together allow us to construct individual models of situations. Philosophers of science tend to agree that these elements are

[10] There has also been some work on developing formal models of ethical theories, such as consequentialism and deontology. This is clearly a normative modeling enterprise as well. See, for instance, Oddie and Milne (1991) and Colyvan, Cox, and Steele (2010).

[11] For further details beyond the brief discussion of this paragraph, see Titelbaum (2013, §4.2). Colyvan (2013) also offers an excellent discussion (with case studies!) of the different types of idealizations in normative models.

[12] I said that the social sciences engage in descriptive modeling; critics of the prominence of rational actors in economic theory may debate whether economics counts.

all required for a modeling enterprise,[13] while disagreeing about which of them should be called the "model". (Sometimes what I've called a "framework" is a "model"; sometimes the combination of framework and interpretation is the "model", etc.) There's also terminological variation among working modelers. But nothing of deep philosophical significance hangs on where we apply the word "model", as long as all three elements are involved under some name or other.

Philosophers who study modeling have rightly emphasized the need for interpretations. Uninterpreted, a formal model is simply a bundle of abstract entities; it neither represents anything in the world nor says anything about it. Yet philosophers have also underestimated the complexity of interpretations. Interpretation is a two-way street: we need to know not only how features of the world should be captured in features of our models, but also how to read off features of models as conclusions about the world. These are *separate operations*; it's important to keep them distinct and identify principles for each.

For example, the Lotka-Volterra modeling framework is standardly interpreted to provide models of the interactions of predator and prey populations. To build a model of a particular ecosystem using this framework, we have to know how to set the model's parameters to match that system. Here the framework is typically interpreted by saying "P represents the size of the predator population," "V represents the size of the prey population", and the parameters r, m, a, and b represent quantities such as birth and death rates, etc. Given these pairings between numerical values in the model and quantities in the world, we can look to the target system, determine the relevant quantities, and set the parameters to build a model of that system.

Yet so far we have only told one portion of the interpretive story.[14] Suppose we build our model by setting the birth rate, death rate, etc. and inputting the predator and prey populations at a specific time. We then use the model's differential equations to calculate predictions for those populations over time. What happens when the model outputs a P-value of, say, 211.47 for some specific time? How are we to read this prediction?

It's easy to address this concern with an offhand remark like "Modelers know which features of their models to take seriously" or "scientific models are only approximations."[15] Such responses make it sound like the model is actually predicting that there will be 211.47 predators at the given time, and we need to be wise enough not to listen to everything the model is saying. Yet sometimes when a model outputs a non-integer value for some variable corresponding to an integer real-life quantity, that value is an *expectation* for the quantity. Expectations are not approximations, nor are they misled predictions that some integer quantity will come to take on non-integer values. And more complex models, such as statistical models, may output a confidence interval or a probabilistic distribution over some

[13] See Weisberg (2013, p. 15, note 3) for references to philosophers of science who have emphasized the combination of structure and interpretation in modeling.

[14] Not to suggest that the part we've already told is always easy. The extended literature on representation theorems in decision theory (Fishburn 1981) shows how difficult it can be just to get an adequate representation of an agent's preferences within a formal belief-desire model.

[15] Cf. Williamson (2017) on "impossible models".

variable. While a model may have been built by inputting a real-world quantity as the initial value of some variable, such outputs make it clear that the model is not just making a point-valued prediction for that quantity.

It's insufficient to interpret a model by saying something like "P represents the size of the predator population."[16] Instead, we should first say that the model is built by setting P equal to the predator population at some time. Then, once some calculations have been made using the model's differential equations, we should provide a separate set of instructions for interpreting the P-value that results. It might turn out that we want to say, "The final P-value represents the model's prediction for the number of predators at such-and-such time," but then again we might instead say, "The model predicts that in the long run, predator-prey systems with such-and-such initial configuration will *average* a predator population of P at such-and-such time," or any number of other things. We need instructions for representing the world in the model, but also instructions for reading out what the model says about the world.

How does this work for normative modeling? For our first example of a normative modeling framework, we'll take the AGM approach to belief revision (Alchourrón, Gärdenfors, and Makinson 1985). The AGM modeling framework is often used to model rational constraints on the changes in an agent's beliefs over time. To build an AGM model of a particular agent, we represent the contents of the agent's beliefs at some initial time as a set of sentences in a formal language. We then represent in the model that the agent is going to be gaining or losing particular beliefs. Both the representation of the agent's initial beliefs and the representation of the beliefs gained or lost can be thought of as setting the parameters of the model.

Suppose that in a particular AGM model, the set K representing the agent's initial beliefs contains sentence p. Suppose further that some time later the agent loses the belief whose content is represented by p (perhaps the agent's reasons for that belief are undermined). The AGM framework provides a set of postulates, common to every model built using that framework, for constructing a new sentence set K÷p. K÷p helps us understand what the agent's beliefs should look like after she loses p. Notice that generating K÷p won't just be a matter of removing p from K; we also want, for instance, to remove any conjunction in K of which p is a conjunct. Alchourrón, Gärdenfors, and Makinson's core contribution was to provide a particular formal operation known as "partial meet contraction" for constructing a plausible K÷p. (The details of that operation are unnecessary for our purposes.)

So far I've mentioned some elements of the AGM modeling framework (the postulates for partial meet contraction) and how to represent features of the world (an agent's initial beliefs and belief-change experiences) in building a model. But we still need to know how to interpret a model's results—how should we understand the sentence set K÷p? One possible interpretation is that if the agent is rational, after losing her belief in p she will possess beliefs corresponding to each sentence in K÷p. Yet this is fairly implausible, for the AGM postulates close K÷p under logical

[16] Perhaps the mistake results from taking the relation between numerical values in a model and quantities in the world to be too much like representational relations between linguistic entities and objects/properties in the world.

consequence. This makes the proposed interpretation too cognitively demanding on agents, for Harmanian “clutter avoidance” reasons (Harman 1986).[17] Levi (1991) suggested instead that we interpret the output sentence set as representing the beliefs to which an agent is *committed* at the later time. Perhaps, though, we don’t even want to say that an agent is committed to various obscure logical consequences of what she already believes, propositions that could never be of any practical use to her. We might then say that while the agent is under no obligation to adopt attitudes towards the propositions represented in K÷p, if she *does* adopt an attitude towards such a proposition, the attitude rationally required of her is belief.[18]

I don’t want to advocate any of these interpretations in particular. I simply want to note that it's not enough to say, "The sentence set represents the agent's beliefs", and also that moving to a more precise interpretation offers added flexibility in how we might understand what our models are saying. Knee-jerk objections to a modeling framework can be avoided by specifying more clearly how its models are to be read.

Given a modeling framework and an interpretation, we can test predictions made by the framework’s models against data. While frameworks make contact with the world (so to speak) through their individual models, it is ultimately the combination of framework and interpretation that we’re testing by comparing those models to data. Going back to the Lotka-Volterra example: Faced with a particular predator-prey ecosystem, we build a Lotka-Volterra model by setting the parameters and initial variable values in the differential equations. Given the interpretation we’re using, the model makes predictions about how the predator and prey populations will behave going forward. We then compare those predictions with data about the real-world populations over time. If there’s a significant mismatch, we take this as a lesson *about the interpreted Lotka-Volterra framework*, not just about that particular model. A bad enough mismatch may lead us not only to abandon this particular model, but also to question whether we should apply the Lotka-Volterra equations under this interpretation to other populations as well. Though this oversimplifies the relation of models to their frameworks, one way to think about what’s happened is that the Lotka-Volterra *framework*, under a particular interpretation, has made a prediction of the form, “If you have a system with such-and-such parameter settings and initial populations, after so much time this-and-that will happen with the populations.” The data mismatch shows that this conditional prediction is false, and may call the entire interpreted framework into question.[19]

[17] Yap (2014) tries to deal with such problems by reading AGM models as highly idealized *descriptive* models. I will not pursue that approach here.

[18] Notice that this proposal is different from saying that the agent is rationally permitted to believe the propositions represented in K÷p. Saying that the agent is permitted to believe such a proposition (call it q) is consistent with saying that she’s also permitted to disbelieve q, but that’s being ruled out here. Also, the agent might not *currently* be permitted to believe q, because there might be other things rationality requires her to accomplish before she’s permitted to do that (such as carefully considering q, noticing its connections to other propositions in K÷p, etc.).

[19] More on this when we discuss domains of applicability.

Return now to AGM as a modeling framework. We build a model by representing a particular agent's initial beliefs as a sentence set, then represent her loss of a specific belief. Applying the AGM postulates, we determine the features of K÷p, then apply our chosen interpretation to draw normative conclusions. Perhaps the model indicates that after the loss of belief, rationality requires the agent to believe such-and-such propositions. Now suppose that this prediction mismatches the data: It's not the case that rationality requires the specified beliefs.[20] This calls the combination of AGM and our interpretation into question. At that point we have two options. We may leave the formal framework alone, and modify our interpretation—as Levi did with AGM. But if the mismatch is bad enough, or we cannot find an interpretation that solves the problem, we may abandon or modify the formal framework itself. For example, in light of various putative counterexamples that arose after AGM's original publication, a number of philosophers have proposed alternatives to the framework's partial meet contraction operation.

4. Advantages of modeling

I now want to describe some advantages of applying a modeling methodology in the course of a normative inquiry. In fact, all of these strike me as general advantages of formal modeling. Many of them will be familiar from the setting of descriptive modeling; I want to highlight how they can be advantages in normative settings as well.

• *Clear separation of model from target*. On a formal modeling methodology, equations and strings of symbols describe a model, which is an abstract structure. The model is then related via explicit principles to real-world phenomena and/or norms applying to those real-world phenomena. There is no conflating what's in the model with what's in the world.[21]

This clear separation has a number of advantages—in fact, many of the advantages I'll list below rely crucially on the separation. We've already seen one example of such an advantage in our discussion of the deductive closure of sentence sets in AGM models. For a number of reasons it's formally convenient to close AGM sentence sets under logical entailment. But as we saw earlier, *not everything in the model needs to be in the agent*. The presence of a particular sentence in an AGM set *might* indicate that an agent is rationally required to possess a particular belief, or at least be committed to that belief. But this all depends on our interpretive principles; we *need* not read the presence of a particular element in a formal AGM model as requiring anything of an agent out in the world. The model is an abstract structure

[20] Notice that the issue here is whether *rationality requires* particular beliefs, not what beliefs actual agents tend to have after losing particular bits of information. That's what makes this a normative modeling endeavor rather than a descriptive one.

[21] Compare Morgan and Morrison (1999, p. 11): "What it means for a model to function autonomously is to function like a tool or instrument. By its nature, an instrument or tool is independent of the thing it operates on, but it connects with it in some way."

with various mathematical features; the model doesn't require those features in agents unless our interpretation says so.

This is just one example of the flexibility gained by clearly distinguishing between a formal model and its real-world target. When an author writes down normative principles directly, those principles apply immediately to the given target system. But a modeling methodology has more moving parts—formal models, equations elaborating their framework, the target phenomena, etc.—each of which must be connected to the next. This adds some amount of methodological complexity, but also provides extra wiggle room to make adjustments. In the AGM case, we can select simple and powerful formal tools for use within the models, then compensate for any oddities produced by adjusting our interpretation.

• *Choices must be made explicit*. In specifying the abstract structure that is a formal model, the modeler must make explicit particular choices that are easy to elide when operating in a less formal fashion. I'll illustrate this point with an example from a normative modeling framework I've worked with extensively, Bayesian epistemology.

Understood as a modeling framework, Bayesian epistemology models rational constraints on an agent's degrees of belief—both at a given time, and as those degrees of belief (or "credences") change over time. To build a Bayesian model, we first select a formal language of sentences to represent potential credence targets. We then construct a "credence function" to represent the agent's degrees of belief at some initial time; each degree of belief is represented by a real-number value assigned to a sentence in the model's language.[22] Finally, if the agent gains evidence after the initial time, that evidence is represented as a set of sentences in the model's language.

The Bayesian modeling framework provides a number of mathematical constraints on credence functions. Some of these constraints—to which we will return later—are inspired by Kolmogorov's (1933/1950) probability axioms. But for now we will focus on another Bayesian constraint, known as Updating by Conditionalization. Conditionalization takes as inputs the credence function representing the agent's initial degrees of belief and the set of sentences representing the evidence the agent gains, and outputs a new credence function. This new credence function is usually interpreted as indicating rational constraints on the degrees of belief the agent assigns after assimilating her new evidence.

Most Bayesian epistemologists don't think of themselves as modelers, so most Bayesians don't follow the methodical practice I've just described. Faced with a situation in which an agent assigns some degrees of belief, a Bayesian will just start writing down equations describing those degrees of belief (where these equations are best understood via something like the Abbreviated Description interpretation). The Bayesian will then apply Conditionalization to generate

[22] Credence functions in Bayesian models are sometimes assigned over propositional languages rather than sentential. Switching from sentences to propositions would require me to present some of the material below in a slightly different way, but ultimately the switch wouldn't make any significant difference.

credences for some later time, and her analysis is off and running. Rarely does a Bayesian pause before writing equations to first define a language over which the credence functions will be assigned, for example by specifying its complete set of atomic sentences.

Why should this matter? Unlike deductive relations, probabilistic relations can be nonmonotonic. If we find that a deductive relation (say, entailment or inconsistency) holds between two sentences in a language, that relation will still hold when the language is given more fineness of grain (perhaps by adding atomic sentences). But results that hold for credence functions defined over one language may fail when a finer-grained language is used.[23] For example, it's important that the set of evidential sentences to which Conditionalization is applied represents "everything the agent learns" after the initial time. If our language representing potential credence targets is too simple, we may miss some of the (relevant) information the agent learns, because that information is incapable of being represented in the language we've chosen.[24]

A Bayesian pursuing a modeling methodology must define her formal model exactly and completely before applying it to a problem. In the process, she will (among other things) specify the sentential language over which the model's credence functions are assigned. This forces her to make an explicit choice about what to include in the language and what to leave out. This choice may be difficult, and making it judiciously may require substantive philosophical work. (The main threat is that by leaving a significant sentence out of the model's language, the modeler may fail to represent relevant information learned.) But it's important for Bayesians to make these choices, make them up-front, and make them explicitly. For one thing, this bit of methodological hygiene may spark a broader conversation about how such choices should be made.[25] For another, it keeps Bayesians from proceeding without a declared modeling language and making errors they haven't even given themselves the representational resources to identify.

• *Counterexample management through domain of applicability*. Suppose I state a normative principle: rationality requires agents to do such-and-such. Then you produce a counterexample, in which it looks like an agent can fail to do such-and-

[23] A similar concern arises in decision theory, connected to Savage's (1954) discussion of "small world" decision problems. Joyce (1999, p. 72) writes, "Choosing is really a two-stage process in which the agent first refines her view of the decision situation by thinking more carefully about her options and the world's state until she settles on the 'right' problem to solve and then endeavors to select the best available course of action by reflecting on her beliefs and desires in the context of this problem. Decision theorists have concentrated almost exclusively on the second state of this process."

[24] Lest you think this would never actually happen in practice, Bostrom (2007) attacked traditional Bayesian approaches to the Sleeping Beauty Problem by accusing them of overlooking a relevant set of sentences learned by Beauty between two times. Bradley (2010), meanwhile, dispelled a proposed counterexample to Conditionalization by arguing that it represented an agent's learning with too impoverished a modeling language.

[25] In Chapter 8 of Titelbaum (2013), I consider some principles for selecting the language for a Bayesian model, and some principles that interrelate models defined over different languages.

such while remaining perfectly rational. If I agree with your verdict about what's rational in the example, how should I respond?

A number of moves are familiar from the philosophical literature on norms. Most often, the theorist abandons the principle and looks for another one that accommodates the counterexample. Yet other evasive maneuvers are available.[26] Sometimes the normative theorist will keep her principle largely intact, but say that it specifies only a *pro tanto* or *prima facie* reason to behave in a particular fashion—a reason that can be overridden in particular cases. Or the normative theorist will say that her principle holds only *ceteris paribus*, and the cited counterexample is not one in which the relevant things are equal.

How does counterexample management work on a modeling methodology? I said before that a modeler can respond to a mismatch between a model and data she accepts either by making an adjustment in her modeling framework, or by making an adjustment in its interpretation. The former might involve changing the equations and rules that specify the framework, or even trading that framework in for another. What about changing the interpretation? Here's where maneuvers like reinterpreting the model's predictions as *pro tanto* or *ceteris paribus* become available.

Yet there are cases in which such reinterpretations are simply inappropriate. To return to Bayesian epistemology, adopting Updating by Conditionalization famously leads Bayesians to odd consequences. Recall that Conditionalization takes as its inputs an initial credence function and a set of sentences representing evidence the agent gains. In combination with other Bayesian norms, Conditionalization sets the posterior credence of every sentence in the evidence set to 1—the maximal credence value on a Bayesian approach, usually interpreted to represent absolute certainty. Moreover, Conditionalization is often iterated when an agent gains multiple pieces of information at different times. Suppose we apply Conditionalization to an initial credence function to generate a later credence function. Then we apply Conditionalization to that later function to generate yet a further function when the agent learns something new again. As we repeat this process, any sentence that gets sent to credence 1 at any point will retain that credence through future Conditionalizations. Bayesian epistemology makes it look like rationality requires an agent to become certain of any piece of evidence she gains, then retain that certainty ever afterwards.

Since most epistemologists have now abandoned foundationalist epistemologies based on indubitable phenomenology, Bayesians concede that real learning experiences are rarely like this, if ever. Bayesians have explored alternative updating norms, such Richard Jeffrey's (1965) "probability kinematics" (now universally known as "Jeffrey Conditionalization"), which don't require evidence gained to become certain. Yet at the same time we Bayesians keep teaching Conditionalization to our students, and keep applying it to solve all sorts of problems in epistemology and confirmation theory. It can't be that we interpret Updating by Conditionalization as a norm providing defeasible *pro tanto* or *prima*

[26] Much of the next few pages could be helpfully compared with the insightful discussion of how mathematicians respond to counterexamples to their theorems in Lakatos (1976).

facie reasons. It's not as if Conditionalization is telling you that you have *some reason* to alter your credences according to a particular mathematical rule, while you may have some reason (or more reason) to do otherwise. Moreover, Conditionalization is not a *ceteris paribus* rule.[27] So what makes us comfortable continuing to apply it?

Here's one proposal for understanding Conditionalization that would square with typical Bayesian practice: We take Conditionalization to be one of the elements in a particular Bayesian modeling framework. Like any modeling framework, the Bayesian framework requires an interpretation. We've already seen that an interpretation contains instructions for representing aspects of the target in a model, and instructions for reading the model's outputs back out to the target. But an interpretation should also include a domain of applicability.

When discussing which cases to model within a particular modeling framework, philosophers of science often assume that a framework applies to any situation that can be represented in one of its models.[28] Typically, though, a modeling framework is intended to be applied only to a proper subset of the targets to which it could be applied. The parameters of a Lotka-Volterra model could be set to match the properties of any two populations, but the Lotka-Volterra framework is meant to apply only to populations in a predator-prey relationship. The domain of applicability in a framework's interpretation specifies which of the things the framework *could* be applied to it *should* be applied to.

For example, a Bayesian modeling framework that includes Conditionalization as one of its rules *could* be applied to a situation in which an agent has a learning experience but nevertheless gains no certainties. Such an agent's degrees of belief over time *could* be represented in a model built using this framework. The resulting model would yield bizarre, implausible verdicts about what rationality requires in that situation. Yet on a sophisticated, modern interpretation of the Bayesian framework in question, the situation isn't a counterexample to the framework, because *the framework isn't intended to be applied to such a situation*. Every contemporary Bayesian understands that a Conditionalization-based approach is appropriate only when an agent learns by gaining certainties, and only when those certainties are to be retained. As I noted earlier, such cases may be highly idealized and rare in real life. But these idealized cases in which new evidence is treated as certain may be, and historically have been, very useful in the analysis of various pieces of scientific and decision-theoretic reasoning. Moreover, cases that don't fit this mold simply aren't part of the domain

[27] Conditionalization in particular is ill-suited to play the role of a *ceteris paribus* rule. That's because Conditionalization is often applied as part of a Bayesian apparatus for determining which pieces of evidence are relevant to a particular hypothesis and which are not. In other words, Conditionalization is part of the tool we use to determine when "all else is equal". The application of Conditionalization therefore cannot rely on an antecedent capability to determine whether all other things are equal or not. (For more discussion and a pointed example, see Titelbaum (2015).)

[28] cf. Morgan and Morrison (1999, p. 20).

of applicability of the Bayesian framework in question. So they constitute no counterexample to that framework, properly understood.[29]

On first encounter this maneuver may appear *ad hoc*. But it's not *ad hoc* as long as it's not performed in an *ad hoc* fashion. When a modeler constructs a modeling framework, she must specify its domain of applicability—the set of targets for which the framework's models are supposed to yield reliable verdicts—as part of the framework's interpretation. This should be done in a principled fashion, by specifying *classes* of cases to which the framework either does or doesn't apply. Ideally the boundaries of the framework's domain will be *justified* or at least *explained* by reference to features of the framework.

On a modeling methodology, there are two main kinds of threats to a modeling framework: First, someone might produce a genuine counterexample—a case that lies *within the framework's specified domain of applicability* yet for which the framework makes inaccurate predictions. In the face of such an example, either the framework or its interpretation must be altered (perhaps by redrawing the boundaries of the intended domain). On the other hand, a case that is capable of being modeled in the framework yet which lies outside the specified domain of applicability poses no threat of potential counterexample.[30]

The second kind of threat to a modeling framework is when someone produces an alternate framework that gets all the same cases right, plus some more besides. A new framework whose domain is a superset of the old's threatens to supersede the old framework on usefulness grounds. Here I am largely in agreement with Williamson when he writes,

> *Counterexamples* play a much smaller role in a model-building enterprise than they do in traditional philosophy.... What defeats a model is not a counterexample but a *better model*, one that retains its predecessor's successes while adding some more of its own.... If epistemologists and other philosophers start aiming to build good models rather than provide exceptionless analyses, different forms of criticism become appropriate. (Williamson 2017, emphases in original)[31]

[29] A similar point could be made about Jeffrey Conditionalization, intended as a replacement for or generalization of the Updating by Conditionalization. Jeffrey himself (1965) recognized that his updating rule was appropriate only when a particular Rigidity condition was met. Later authors such as van Fraassen (1981) produced learning examples that seem to violate Rigidity and therefore lie outside a Jeffrey Conditionalization framework's domain of applicability.

[30] Arntzenius (2003) attacked Updating by Conditionalization using counterexamples involving memory loss—the point being that an agent who forgets may fail to keep earlier evidence certain. Schervish, Seidenfeld, and Kadane responded, "We do not agree with Arntzenius that, in the examples in his article, [Conditionalization] is subject to new restrictions or limitations beyond what is already assumed as familiar in problems of stochastic prediction.... The literature on stochastic prediction relies on [the following assumption] regarding states of information and the temporal variables that index them: When $t_2>t_1$ are two fixed times, then the information the agent has at t_2 includes all the information that she or he had at time t_1." (2004, pp. 315–6)

[31] Williamson's talk here of "defeat" is a bit stronger than I would be willing to go. While an old model may be superseded by a new model with, say, a strictly larger domain of applicability, it might nevertheless be good to keep the old model around. For instance, for the cases lying within both models' domains of applicability, the old model might be more computationally efficient than the new.

• *Treating a model as a unit*. One might respond to this domains-of-applicability approach by saying that the same effect could be achieved without a modeling methodology; one need only articulate individual normative principles as conditionals. For instance, Conditionalization could become: "As long as an agent learns by gaining certainties and keeping them, rationality requires her to update by the following mathematical rule...." Yet on a modeling methodology, the rules or equations used to specify a modeling framework are not to be evaluated singly. Instead, the entire set of such rules is to be taken as a unit, and evaluated by appraising the abstract structures that unit creates.

For example, I mentioned earlier that besides Conditionalization, Bayesians include Kolmogorov's probability axioms among their rational norms. One of these axioms is often stated as follows:

Normality: For any tautology T, cr(T)=1.

On something like the Abbreviated Description interpretation, Normality says that a rational agent assigns every tautology a credence of 1. Besides the kinds of clutter avoidance objections we considered earlier for AGM, this supposed norm is open to the following objection: rationality is supposed to be about consistency relations *among* attitudes, but this norm places a requirement on *single* attitudes (the agent's credences in tautologies), one at a time.

From the perspective of a modeling methodology, this is the wrong way to understand Normality. A modeler would use Kolmogorov's three axioms to help specify a particular kind of abstract object: a probability distribution over a language of sentences. The interesting question to ask then is whether probability distributions are the right formal tool for modeling a particular kind of situation. In this context it's no good to evaluate Kolmogorov's axioms one at a time, absent their connections and interrelations to the other axioms.

This point is reinforced by the fact that, as any mathematician knows, a particular abstract structure may be axiomatized in many different ways. Instead of selecting Normality as an axiom, we could have adopted a rule that says, "For any sentences P and Q, if P entails Q then Q's credence is at least as great as P's." Adopting this rule has the effect of setting all tautologies to an equal, maximal credence value (since each tautology is entailed by every other sentence, including the other tautologies). So this rule accomplishes the same effects as Normality.[32] No one would complain that *this* rule fails to express a relation among multiple attitudes. But this just shows how misplaced the original complaint was; a substantive objection to our framework shouldn't be avoidable simply by presenting that framework with an alternative yet extensionally equivalent axiomatization.

In the end, what should get evaluated (and what *does* get evaluated by confrontation with data) is an entire modeling framework, with all its rules

[32] The selection of 1 as the maximal credence value assigned to all tautologies is strictly conventional.

contributing to an organic whole.[33] It is the framework that receives an interpretation, and it is at the level of the whole framework (rather than at the level of individual rules) that we consider domains of applicability. For the Bayesian framework, Normality is one element of *one specification* of a particular abstract mathematical object; the resulting object is then interpreted as indicating norms of rational consistency among attitudes. To a modeler, Normality was never meant to stand alone as a rational norm. Its significance, and its advantages, can be seen only in light of the other Bayesian rules, and in light of the effects it produces in concert with those rules.

Return to Conditionalization, and to the proposal to restate it with an antecedent barring its application to particular kinds of situations. The trouble with this proposal is that Conditionalization *by itself* does not create problems in those situations. People often *say* that Conditionalization generates and retains certainties, but it doesn't do that on its own: Conditionalization has these effects only in combination with Kolmogorov's probability axioms and another Bayesian rule known as the Ratio Formula. So the entity that should be restricted to apply only to situations in which agents gain and retain certainties is the entire modeling framework—combining Conditionalization, Kolmogorov, and the Ratio Formula. The entire modeling framework is the proper object of interpretation, and part of that interpretation is the principled specification of a domain of applicability.

• *Multiple interpretations of the same framework*. Finally, once a modeling framework is understood as a template for building formal models, and interpreting that framework is defined as a separate task from writing down rules that specify the framework itself, we open up the possibility of offering multiple interpretations of the same modeling framework. This situation is familiar from the sciences. For example, the mathematical structures of what we still call the "fluid dynamics" framework are now used for modeling much more than just fluids moving around.

Nowadays when the principles of AGM are described, they are often presented as "If an agent believes such-and-such, then comes to learn such-and-such, she should…." But in a retrospective on the development of AGM, Gärdenfors writes,

> I came into the models of AGM from philosophy of science and counterfactual reasoning. The motivation of Alchourrón and Makinson was originally derived from legal theory. It came as a surprise to us that the field where the AGM theory had the strongest immediate impact was in computer science as a theoretical foundation for database updates. (Gärdenfors 2011, p. 118)[34]

Later in the same piece, Gärdenfors discusses adopting principles of belief revision into principles for non-monotonic reasoning. There's no reason why the "theories" that appear in AGM models must be interpreted as sets of beliefs, rather than sets of norms in a legal code or items in a database. Separating abstract structures from

[33] Cf. Colyvan: "As with empirical models we are interested in whether a given model—as a whole—is adequate for the purpose for which it was designed." (2013, p. 1348)

[34] Notice also that the very first paragraph of the famous (1985) AGM paper contains a reference to the process "known among legal theorists as the *derogation* of *x* from *A*."

their interpretations allows the same structures to be used in multiple applications.[35] And there's no reason a particular abstract structure must be used *either* normatively *or* descriptively but not both; Bayesian and decision-theoretic formalisms have been applied in psychology, cognitive science, economics, etc.

5. Concerns about normative modeling

I will now examine some concerns that might arise about applying a modeling methodology to normative inquiries.

• *Systematizability of the domain.* The idea that we can identify a domain of applicability, fit a formal modeling framework to truths in that domain, then rely on that framework to predict further normative truths within the domain, seems to make assumptions about the nature of the normative. Those assumptions are tricky to state precisely without prejudging the issue, so let me instead identify them by describing some types of theorists who would resist them.

First, I imagine that normative particularists (e.g., Dancy 2013) will object to a modeling methodology. They will wonder why we assume that the normative truths in distinct situations fit into any sort of pattern, such that a framework designed to fit some cases will be able to predict truths in others. Of course, the modeling methodology I propose is not the only normative approach subject to particularist complaint. In fact, normative modeling may not be as objectionable to the particularist as the typical method of normative theorizing by laying out principles. That method usually aims for principles that are fully general; hence its aggressive reaction to proposed counterexamples. That such principles are available assumes the denial of particularism. Normative modeling, on the other hand, attempts to systematize limited swaths of the normative domain, identified as the domains of applicability of particular modeling frameworks. There need be no assumption that *fully* general principles are available. 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.

Second, some non-particularists may acknowledge patterns among the normative truths, but deny that these patterns can be represented in *formal* structures. For instance, Goodman famously concluded from his discussion of the grue paradox that, "Lawlike or projectible hypotheses cannot be distinguished on any merely syntactical grounds or even on the ground that these hypotheses are somehow purely general in meaning." (1955, p. 83) Goodman meant to bury Carnap's and Hempel's project of capturing confirmation (a normative notion) in a

[35] Compare the notion of "re-semantification" in Dutilh Novaes (2012, Section 6.1.2), and also Weisberg (2013, p. 77): "The same structure can be given an entirely new construal when a model is borrowed from another area or problem."

formal construct. Perhaps some normative domains are systematizable, but not by formal means.

I tend to disagree with Goodman about the particulars—problems like grue that look like formal headaches usually turn on a deep issue having nothing to do with formalization (Titelbaum 2010). But I certainly can't argue against his conclusion here, or support any claim to the effect that all generality supports formal representation. I will simply say that in this piece I am trying to describe formal normative modeling and contrast it with other methodologies *formal* normative theorists might apply. If the underlying normative domains are not systematizable by any formal means, all of these methodologies are in equal trouble.

• *Whence the "data"?* Earlier I distinguished normative from descriptive models in terms of the data they attempt to fit. Normative models attempt to fit normative data, such as prescriptive or evaluative facts. One might wonder where a modeler is to acquire such "data".[36] Is normative modeling just a cover for the modeler's codifying her own intuitions?

The origin of the data is an important concern for the model*er*, but is not a concern for the model*ing* process itself. In the sciences, gathering reliable data is a crucial project, but the process of building a model from the data is usually separate from that project.[37] The modeling process asks what predictions fit best with a given set of data; even if the factivity of the data is later challenged, the model's results about what predictions best fit the data still stand. Of course, fitting certain batches of data with a model sometimes proves so difficult that it motivates the modeler to reexamine the data. (Perhaps the modeling effort has made specific subsets of the data particularly suspect.) But usually a modeling enterprise ventures forward from the data as a given.

Normative modeling requires a set of normative facts taken as given. Those facts may have been provided by intuition, but they may also have been provided by rigorous argumentation. They may even have resulted from a previous modeling exercise![38] Certainly the trustworthiness of any predictions produced by the modeling depends on the trustworthiness of this data. (Again, we can roughly think of the modeling framework as yielding conditional predictions: "If such-and-such are normative truths, then so-and-so are as well.") But as long as one grants that there are normative facts at all, and that some of them can (somehow) be known,

[36] Compare's Colyvan's concerns about "empirically testing" normative models at his (2013, pp. 1347ff). With respect to normative models one might also worry whether there are any normative truths to serve as data at all. But such metanormative skepticism (perhaps in the mode of Mackie (1977)) will be a challenge to any of the methodologies of normative theorizing with which a modeling approach might be compared, so it is not a *particular* challenge to our approach.

[37] Which is not to say that modeling is always *posterior to* data collection. Good modeling may send us back to our data-gathering mechanisms to ask further questions. My point here is that the process of modeling itself is usually distinct from whatever tools we are using for data collection.

[38] This phenomenon is hardly unique to normative modeling. For example, Edwards (2010) details how the data fed into global climate prediction models are often the output of more local models applied to unreliable instrumentation or incomplete records from the past.

normative modeling is an appealing candidate methodology for extending that normative knowledge base.

• *What modeling achieves.* Beneath the two previous concerns I sometimes detect a broader suspicion: Even if a modeling exercise is successful, what will it have achieved? Models may track local generalities over limited domains of applicability, but we wanted the true, fully general normative principles. Modeling may take us from normative givens to further conclusions, but we wanted absolute normative truth derived from *no* assumptions—to trace the normative back to its foundations. Modeling may produce formal structures fitting a set of data, but we wanted *understanding*.

My first response to these challenges is that I don't picture formal normative modeling as the only methodology we should use to investigate the normative, or even the only *formal* methodology we should use. Formal normative modeling is a tool—one of many tools in our philosophical toolbox—and like any tool it is appropriate for only certain purposes.

So what *can* formal normative modeling achieve for us? First, it can give us answers to specific normative questions. If we don't know what would be required, or what would be rational, in a particular situation, a normative model fit to other known normative facts can give us an answer. Second, simply providing a formal framework that fits normative facts over a limited domain can reveal patterns that aid our understanding. And third, formal normative modeling can be used in tandem with other methodologies to pursue the broader goals just described. By moving from narrow models to models with broader and broader domains of applicability, perhaps we can approach the fully general principles in some normative area (if there are any). Nowadays we view Newtonian mechanics as a formalism useful only over a limited domain, but no one would deny that it has aided our understanding of the natural world and was a crucial building block towards the theories that supplanted it.

Let me give one example of a positive product from a formal exercise in a normative domain: Bayesian epistemology can be used not only to generate new credence functions from old, but also to assess how strongly a piece of evidence confirms a hypothesis relative to a particular credence distribution. Hempel's famous Paradox of the Ravens (Hempel 1945) asks why observing a black raven confirms the hypothesis that all ravens are black more than observing a non-black non-raven does.[39] Fitelson and Hawthorne (2010) used Bayesian mathematics to show that a rational agent will take a black raven to confirm the ravens hypothesis more strongly than a non-black non-raven does when the following two conditions are met: (1) the agent takes her sampling process to be such that the ratio of non-black objects to ravens it produces will tend to be greater than 1; and (2) learning

[39] The crux of the problem being that the ravens hypothesis that all ravens are black is logically equivalent to the hypothesis that all non-black things are non-ravens; the latter seems to be confirmed by any non-black non-raven; and by some sort of transitivity of confirmation it therefore seems that a non-black non-raven should confirm that all ravens are black.

that the ravens hypothesis was true would not dramatically increase this ratio for the agent.

This is a novel prediction. I take it that these sufficient conditions do not just spring into your mind upon contemplation of the problem; historically, no one had suggested these two particular conditions as sufficient until Fitelson and Hawthorne performed their formal analysis. Once uncovered, the conditions are borne out by various considerations. For instance, it's plausible that as we wander around the world and encounter objects at random, we take ourselves to be implementing a sampling process that satisfies the two conditions. And in these circumstances, to the extent we're rational we take our encounters with black ravens to be better news for the ravens hypothesis than, say, our run-ins with red robins. On the other hand, if I were wandering through the Hall of Atypically-Colored Birds (featuring birds of a different color than the majority of their species brethren), encountering a black raven would be much *worse* news for the ravens hypothesis than encountering a red robin. Yet sampling from the Hall of Atypically-Colored Birds does not meet the two conditions above.[40]

These considerations support the sufficiency of Fitelson and Hawthorne's conditions. But by far the best *argument* for that sufficiency comes from the Bayesian analysis itself. The Bayesian formalism has provided independently plausible results in other contexts, so we believe the normative prediction it makes about how a rational agent should see the confirmational landscape in the ravens case.[41]

6. Conclusion

As I indicated at the outset, this essay has been generally exploratory. I have tried to sketch a formal modeling methodology for normative inquiries, indicate its advantages, and assess its potential drawbacks. A clear area for further research would be to compare this methodology to other methodologies that might be applied to normative formal work, such as reflective equilibrium (Rawls 1971) and Carnapian explication (Carnap 1950).

Yet to make a fair comparison, each of those methodologies would need to be developed in slightly new directions. Rawls' reflective equilibrium approach was inspired by Goodman (1955). But while Goodman made his brief remarks in the context of assessing formal systems, hardly any of the literature on reflective

[40] A bit more explanation why the conditions are and are not met in the two sampling contexts: In everyday sampling we expect there to be many more non-black things around us than ravens. Moreover, learning that all ravens were black would probably tend to increase the number of black things we expected to see, not increase the count of *non*-black items. So the crucial ratio would not increase (or at least not *dramatically* increase) were we to gain this information. On the other hand, learning that all ravens are black would make us surprised to find any ravens at all in the Hall of Atypically-Colored Birds (because if all ravens are black, there *aren't* any atypically-colored ravens). So in that sampling context the ratio does dramatically increase (because its denominator tends towards zero), and the second condition is violated.

[41] While Fitelson and Hawthorne certainly use formal methods, they don't present their results explicitly in the context of a modeling methodology. For novel Bayesian results clearly established using formal models, see Titelbaum (2013, esp. Ch. 6 and 11).

equilibrium that followed Rawls considered it as a methodology for formal work. Carnap, meanwhile, articulated explication as a method for understanding concepts. While some of those concepts were normative, applying explication to a variety of normative domains would require stretching its targets beyond just the conceptual.

Hopefully in the future these methodologies (and others) will be adapted to formal work on norms. At that point we will be able to compare them fairly to normative modeling, and see how multiple methodologies might work in tandem towards greater philosophical goals. For instance, within a broader reflective equilibrium process, a modeling methodology might be used locally to test the consequences of embracing particular considered judgments about cases. But for now, such possibilities remain largely speculative.[42]

[42] I am grateful to Joachim Horvath, Catarina Dutilh-Novaes , and Mark Colyvan for helpful comments on earlier drafts and for bringing important references to my attention. Thanks also to an audience at "2019 Interdisciplinary Colloquium on Probability Theory: Philosophy, Physics, and Mathematics at the Crossroads" in São Paulo, Brazil. Finally, I am grateful to the University of Wisconsin-Madison and to the Australian National University, at which I completed this essay during a Visiting Fellowship.

Bibliography

Alchourrón, C.E., P. Gärdenfors, and D. Makinson (1985). On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic 50*, 510–30.

Arntzenius, F. (2003). Some problems for conditionalization and reflection. *The Journal of Philosophy 100*, 356–70.

Beall, J.C. and G. Restall (2006). *Logical Pluralism*. Oxford: Oxford University Press.

Bostrom, N. (2007). Sleeping Beauty and self-location: A hybrid model. *Synthese 157,* 59–78.

Bradley, D. (2010). Conditionalization and belief *de se*. *Dialectica 64*, 247–50.

Carnap, R. (1950) *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Colyvan, M. (2013). Idealisations in normative models. *Synthese 190*, 1337–1350.

Colyvan, M., D. Cox, and K. Steele (2010). Modelling the moral dimension of decisions. *Noûs 44*, 503–29.

Dancy, J. (2013). Moral particularism. In Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy (Fall 2013 Edition)*. URL = <http://plato.stanford.edu/archives/fall2013/entries/moral-particularism/>.

Dutilh Novaes, C. (2012). *Formal Languages in Logic: A Philosophical and Cognitive Analysis*. Cambridge: Cambridge University Press.

Edwards, P.N. (2010). *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*. Cambridge, MA: The MIT Press.

Fishburn, P. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision 13*, 139–99.

Fitelson, B. and J. Hawthorne. (2010). The Wason tasks(s) and the paradox of confirmation. *Philosophical Perspectives 24*, 207–241.

Gärdenfors, P. (2011). Notes on the history of ideas behind AGM. *Journal of Philosophical Logic 40*, 115–120.

Godfrey-Smith, P. (2006). The strategy of model-based science. *Biology and Philosophy 21*, 725–40.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Grice, H.P. (1975). Logic and conversation. In P. Cole and J.L. Morgan (Eds.), *Syntax and Semantics, Volume 3: Speech Acts*, pp. 41–58. Elsevier.

Harman, G. (1986). *Change in View*. Boston: The MIT Press.

Hempel, C.G. (1945) Studies in the logic of confirmation (I). *Mind 54*, 1–26.

Jeffrey, R.C. (1965). *The Logic of Decision*. New York: McGraw-Hill.

Joyce, J.M. (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

Kolmogorov, A.N. (1933/1950). *Foundations of the Theory of Probability*. Translation edited by Nathan Morrison. New York: Chelsea Publishing Company.

Lakatos, I. (1976) *Proofs and Refutations*. J. Worrall and E. Zahar (Eds.). Cambridge: Cambridge University Press.

Levi, I. (1991) *The Fixation of Belief and Its Undoing*. Cambridge, MA: Cambridge University Press.

Lotka, A.J. (1956). *Elements of Mathematical Biology*. New York: Dover.

Mackie, J.L. (1977) *Ethics: Inventing Right and Wrong*. New York: Penguin.

Morgan, M.S. and Morrison, M. (1999). Models as mediating instruments. In M.S. Morgan and M. Morrison (Eds.), *Models as Mediators*, pp. 10–37. Cambridge: Cambridge University Press.

Oddie, G. and P. Milne (1991). Act and value: Expectation and the representability of moral theories. *Theoria 57*, 42–76.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Savage, L.J. (1954). *The Foundations of Statistics.* New York: Wiley.

Schervish, M.J., T. Seidenfeld, and J. Kadane (2004). Stopping to reflect. *The Journal of Philosophy 101*, 315–22.

Stalnaker, R.C. (1978/2002). Assertion. Reprinted in P. Portner and B.H. Partee (Eds.), *Formal Semantics: The Essential Readings*, pp. 147–161. Oxford: Blackwell Publishers Ltd.

Suppes, P. (1960). A comparison of the meaning and use of models in mathematics and the empirical sciences. *Synthese 12*, 287–300.

Titelbaum, M.G. (2010). Not enough there there: Evidence, reasons, and language independence. *Philosophical Perspectives 24*, 477–528.

— (2013). *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford: Oxford University Press.

— (2015). Reply to Kim's "Two versions of Sleeping Beauty". *Erkenntnis 80*, 1237–1243*.*

van Fraassen, B.C. (1981). A problem for relative information minimizers. *British Journal for the Philosophy of Science 32*, 375–9.

Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. *Nature 118,* 558–560.

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

Williamson, T. (2017). Model-building in philosophy. In R. Blackford and D. Broderick (Eds.), *Philosophy's Future: The Problem of Philosophical Progress*, pp. 106–22. Oxford: Wiley Blackwell.

Yap, A. (2014). Idealization, epistemic logic, and epistemology. *Synthese 191*, 3351–3366.