Permissible Idealizations for the Purpose of Prediction

Michael Strevens

Forthcoming, *Studies in History and Philosophy of Science*

Abstract

Every model leaves out or distorts some factors that are causally connected to its target phenomenon—the phenomenon that it seeks to predict or explain. If we want to make predictions, and we want to base decisions on those predictions, what is it safe to omit or to simplify, and what ought a causal model to describe fully and correctly? A schematic answer: the factors that matter are those that make a difference to the target phenomenon. There are several ways to understand differencemaking. This paper advances a view as to which is the most relevant to the forecaster and the decision-maker. It turns out that the right notion of differencemaking for thinking about idealization in prediction is also the right notion for thinking about idealization in explanation; this suggests a carefully circumscribed version of Hempel's famous thesis that there is a symmetry between explanation and prediction.

Acknowledgments

Thanks to conference audiences in Munich (at the Munich-Sydney-Tilburg conference in 2013 and then at a later MCMP workshop on idealization and explanation); Auburn University ("Representation, Idealization, and Explanation in Science"); and the Czech Academy of Sciences ("Idealization Across the Sciences") for for their insightful questions and challenging objections; also to two anonymous referees for encouraging me to clarify and extend my thinking on a number of points.

1. Introduction

Virtually every scientific model is modified, simplified, massaged, somehow bent out of shape. As a consequence, models typically misrepresent the systems that they purport to depict. Such distortions are, when deliberate, called idealizations. (I follow current philosophical practice, then, by using the term *idealization* to refer to modifications of a model that not only withhold truths but purvey falsehoods. For mere withholders, I use the standard term *abstraction*.)

Not every misrepresentation is a legitimate idealization; to count as such, and therefore to transcend mere error, it must display certain qualities. The philosophy of idealization in science is concerned with the nature of those qualities—with the rules or guidelines or tacit conventions governing idealization—and with the ways in which the conventions can be exploited to advance the interests of science, that is, with the scientific functions of idealization.[1]

Both the form and the function of idealizations might be expected to vary with the uses to which idealized models are put. An idealization that is quite acceptable in a predictive model might, for example, be regarded as an ugly flaw when that model is used to explain the very same phenomenon. Or so it seems reasonable to suppose. In past work, I have written about idealization in explanatory models (Strevens 2008, 2017, 2019). This paper is intended to contribute to a discussion of idealization in the course of prediction.

I will have little to say about the benefits, or scientific functions, of predictive idealization; I will rather be concerned with the question of form, that is, the question of which idealizations are and are not permitted in the course of making predictions. That said, my approach is in a certain sense functional.

1. For a range of approaches to the philosophy of idealization, the reader might consult Cartwright (1983); McMullin (1985); Nowak (1992); Mäki (1992); Suárez (1999); Weisberg (2007); Bokulich (2016); Appiah (2017); Elgin (2017); Potochnik (2017), to select invidiously just a small portion of the research on the topic. The present paper, because of its relatively noncommittal attitude to the question of the function of idealization, is compatible with the spirit of much of this work.

One way to determine the conventions of idealization is to look at the scientific literature and to see which idealizations are in fact tolerated or, indeed, encouraged by working scientists. That is an approach I have taken in my work on idealization in explanation. In this paper I take the other way, which is to consider the task at hand—prediction—and to rule out as illegitimate those idealizations that interfere with its successful performance. They may not be the only impermissible idealizations, but they are presumably among them. In what follows, then, I will be asking: which idealizations tend to impair, and which idealizations leave unmolested, a model's predictive prowess? That is the question of *safety* in predictive idealization.

2. The Question of Safety

2.1 Causal Models

I confine myself, in this paper, to predictive models that are causal in character, that is, models that attempt to predict phenomena using a representation of certain causal structures that characteristically produce those phenomena.

The reason is as follows. The need to idealize creates a clear and specific problem for causal predictive modelers. On the one hand, it seems that the most accurate predictions will issue from an accurate representation of causal structure. On the other hand, the very essence of idealization is to represent causal structure inaccurately.

The same is not true of non-causal models; indeed, you might wonder whether idealization raises any interesting philosophical problems for non-causal models at all. A purely phenomenological predictive model—a model that simply tries to replicate patterns observed in nature—can be more or less accurate, more or less approximate, but there is no clear sense in which it can be idealized in the usual sense of that word. I don't claim that other kinds of non-causal model can't be idealized, but I do think that such idealizations will raise problems that are quite different from the problems posed by causal

idealizations, and thus which ought to be dealt with separately.

I must say something, however, about models that "black-box" certain elements of causal structure, representing only the net effects of these elements and none of their workings. Virtually every scientific model contains some black-boxing of this sort. Most microeconomic models, for example, black-box the psychological processes that drive the decision-making of individual economic actors. They tell you, for any given situation, what choice is made, without saying anything about the way in which the decision is reached.

Scientists are nevertheless prone to talk of the idealization of black-boxed processes in a way that implies causal misrepresentation. A microeconomic model, for example, may be said to idealize by falsely representing humans as perfectly rational. In fact, the model's black box does not represent either rationality or irrationality; rather, it represents a situation-to-decision function that is the one that would exist if the thinker were fully rational. It implies full rationality, as it were, without spelling it out.[2] My discussion of and conclusions about predictive idealization will apply not only to explicit causal distortion, but also to those many cases in which black boxes are considered to incorporate implicit idealizations.

2.2 The Predictive Use of Causal Models

A causal model, as I understand it, is a structured set of representations capturing certain aspects of the causal process that produces the "target phenomenon" that is, whatever phenomenon the model is intended to predict (or for an explanatory model, to explain). Such a model will represent, then, initial conditions, boundary conditions, physical, biological, and psychological structures, laws, local regularities, and whatever other elements of reality play a role in getting things caused.

2. And it implies full rationality only in a loose, pragmatic sense: the decision-making black box, for example, might in principle be implemented by a rather irrational mind, say, one that consults an internet psychic who just happens to issue advice that conforms to the economic ideal.

The representational apparatus of a model may take many forms. (Weisberg (2013) surveys some of the possibilities.) For simplicity's sake, I will think of models as consisting of symbols (sentences, equations, and so on) with propositional contents that, by representing the target phenomenon's causal determiners, jointly entail something informative about the target phenomenon itself. What is entailed will depend, of course, on the initial conditions that are "fed into" the model, which will vary with the predictive scenario.

The information provided about the target phenomenon, like the model itself, may take many forms. Some models are deterministic, entailing a specific course of events. Some are stochastic, putting a probability distribution over a range of possibilities. Some are inexact: rather than determining precise probabilities for the possibilities, they ascribe something less well defined: an interval probability ("between 85% and 100%", say), or a qualitative probability ("fairly likely to occur").

This variety in the form and functioning of causal models is matched by a variety in the notion of prediction and, consequently, of predictive success. Even supposing that we limit the discussion to predictions of discrete events, many different kinds of tasks of which causal models are capable count as "making a prediction". A prediction might be a simple yes or no answer to the question, "Will such and such an event occur?" (for example, will Truman be elected president?). It might be the specification of a probability for such an event. It might be the delineation of an interval during which an event can be expected to occur ("There will be a storm some time between 3 PM and 5 PM"), or it might be the specification of the probability of the event's occurring at least once during an interval, or it might be the number of times that the event can be expected to occur during an interval ("Over the course of a typical long, driving lifetime, you should have a total of three to four accidents").[3]

3. https://www.forbes.com/sites/moneybuilder/2011/07/27/how-many-times-will-you-

To allow for these many styles of prediction in an account of idealization would make for a lengthy and involved paper, and would be rather tedious, as no matters of great philosophical interest would arise. Let me therefore restrict my attention to one particular kind of prediction, leaving it to the reader to determine how the resulting treatment of predictive idealization ought to be generalized.

I take as my canonical predictive assignment the task of answering a yes or no question: will an event of a specified kind occur during a specified interval of time? Will there be a storm tonight? Will there be a recession next year? If such an event does occur (at least once) during the interval, say that the target phenomenon occurred. One kind of model suitable for this sort of predictive task is deterministic: it entails, given the relevant initial conditions, either that the target phenomenon will or that it won't occur. Typically, it does so by entailing something more detailed. In the case of the weather, for example, a model might forecast the occurrence of particular storms at particular times. It is natural to declare that the model predicts the target phenomenon—that it gives the answer "yes" to the predictive question—just in case it predicts at least one storm during the specified interval. But it is not compulsory. Given what we know about the weaknesses or the biases of the model, we might, for example, ignore storms predicted on the interval's cusp. Our predictive interpretation of the model appeals if only tacitly, then, to a criterion that turns the model's forecast into a binary, yes/no answer, a criterion concerning which we have some freedom of choice.

Alternatively, a predictive model may be probabilistic, putting a probability distribution over storms at times. Here, too, we appeal at least tacitly to an interpretive criterion to extract a simple "yes" or "no" from the information supplied by the model. The plainest such criterion would say that the answer to the storm question is "yes" just in case the probability distribution imposed by the model, given the initial conditions, yields a probability equal to or

crash-your-car/

greater than one-half that at least one storm occurs during the interval. Again, however, other interpretive criteria might well be chosen on the basis of our background knowledge about the model's quirks.[4]

A prediction is a success if the model says "yes" and the target phenomenon occurs, or if the model says "no" and the target phenomenon does not occur. A good predictive model will generate successful predictions over a wide range of relevant initial conditions. In this paper—again, seeking to simplify where there is little of philosophical interest at stake—I assume that a range of initial conditions is fixed in advance, asking only about the model's predictive power in scenarios contained within that range.

2.3 Towards a Safety Criterion

The purpose of this paper is to pose the question of safety for predictive idealization in causal models: what kinds of idealization can be made without degrading a causal model's predictive power?

The question presupposes that the modeler has a certain predictive goal in mind, since the safety of an idealization will vary with the nature of the goal. It also presupposes that the modeler possesses information about the relevant causal structure that is sufficient, at least in principle, for the building of a model that is capable of some degree of success in attaining that goal (or else the question simply does not arise). What the modeler needs to know is which aspects of the known causal structure simply must be represented accurately in any model that they set out to construct, and which can be omitted or distorted.

The answer I provide will take the form of a criterion that ascertains the safety of an idealization on the grounds of the ultimate determinants of safety,

4. We might also take into account practical matters such as the relative cost of false positives and false negatives. That would, however, complicate the matter of measuring the model's success, adding desiderata above and beyond the very straightforward gauge of success presented in the next paragraph.

that is, by drawing on those features of the situation that ultimately explain why some idealizations are safe and some are not. Such a principle constitutes, as it were, a philosophical theory of safety.

A principle of this sort might be used by the modeler as a method for deciding which idealizations are safe. But in modeling, as in most other endeavors, it is rarely sensible to base decisions on ultimate criteria; typically, it is far more practical to consult some appropriate heuristic. Most modelers seeking answers to questions about safety will find themselves in a situation where various idealizations and their weak points are already well known. An evolutionary biologist modeling gene change in a certain population, for example, will have at their fingertips a range of modeling techniques, some highly idealizing and some less so: models that assume infinitely large populations and models that represent population size accurately; models that assume non-overlapping generations and models that allow overlap; models that build in population structure (e.g., the age of the organisms) with various levels of detail; and so on. They will have some sense of which of these models to use for which predictive questions about which biological scenarios. It is this specialized knowledge that they rely on in putting together their representations of the world; they never need contemplate, let alone put into action, such a thing as a philosophical criterion for safety.

A modeler exploring virgin territory, with no prior knowledge to fall back on, can always answer the safety question from first principles, using the kind of philosophical criterion I attempt to formulate in this paper. But even then, there are alternatives that might have practical advantages. The modeler might, for example, simply proceed by trial and error, trying different combinations of idealizations until they find the one that yields the most reliable predictions.

The vast majority of modelers may never make use of a philosophical safety criterion. Such a thing is, however, of great interest to philosophers of science, because it is what explains the modelers' success: it spells out what it takes for any given heuristic to be reliable—to be responsive in the right

way to the ultimate determinants of safety—and thus what it is for any given application of a heuristic to deliver the correct answers for the right reasons.

A few further comments about the scope and limits of a philosophical safety criterion. First, even in cases where a criterion for safety delivers clear verdicts about what idealizations are feasible, this information might be out of reach for working scientists, especially when modeling systems of great causal complexity. The scientists may lack relevant heuristics. The task of determining safety from first principles may be intractable. And they may lack the computational resources to determine safety easily using trial and error. Such difficulties ought not to be regarded as an objection to the safety criterion in question; they are simply an illustration of the fact that science can be quite hard to do.

Second, a safety criterion says and assumes nothing about the function of idealization; that is, it says nothing about the reasons that you might want to misrepresent a part of a causal process, the outcome of which you are trying to predict. Because this paper is entirely concerned with the question of safety, the question of function will be put to one side. It will be useful, all the same, to articulate one well-known benefit of idealization, just to give a little concreteness to the discussion that follows. I choose perhaps the most uncontroversial example: sometimes we idealize in order to simplify a model, thereby rendering the model easier to analyze or simulate and so making the predictive task easier to accomplish. Ignoring weak forces, small correlations, or minor irregularities can greatly speed calculation; indeed, in some cases, such idealizations are necessary to make calculation tractable at all.

Third, a safety criterion can give only a necessary condition for the permissibility of an idealization, not a condition that is both necessary and sufficient. That is because the advisability of making an idealization depends on other considerations besides safety. We would not, for example, introduce an idealization whose sole effect was to make a model more complicated—an "infelicitous falsehood".

Fourth, an idealization's being unsafe ought not always to prevent its being introduced into a model, as other considerations relevant to idealizing may be in tension with safety, requiring a tradeoff. An unsafe idealization necessarily reduces the predictive reliability of a model, but perhaps only a little. Might we not tolerate, even welcome, a slight drop in accuracy for a sufficiently large gain in simplicity, or some other benefit of idealization? We surely would. Even so, in order to trade off intelligently, we must assess safety. The interest of a criterion that reveals the ultimate grounds of safety is therefore undiminished.

2.4 The Differencemaking/Idealization Thesis

As far as predictive accuracy is concerned—given the way I have set up the predictive task in section 2.2—all that matters is whether a model says "yes" or "no". An idealization can be unsafe, then—it can detract from a model's accuracy—only by affecting the model's answer to the binary question. A causal model delivers this answer on the basis of its representation of the causal process producing the target phenomenon. Some aspects of this "target system" make a difference to whether or not the phenomenon occurs, while some do not. These latter factors, the nondifferencemakers, are causal influences on the phenomenon, and typically affect the way in which the target phenomenon is manifested—the exact timing of a storm, say, or the severity of a recession—but they cannot themselves alter the course of events so that the target phenomenon occurs rather than not occurring, or vice versa. Differencemakers act, in effect, as on/off switches for the target phenomenon in at least some circumstances, whereas nondifferencemakers act as knobs that control the nature of the phenomenon when it does occur but cannot, under any circumstances, divert the course of events so as to turn the phenomenon off or on.

An idealization is a distortion of some element of causal reality. It seems, then, that while distorting a differencemaking causal factor may affect a

model's predictive accuracy, distorting nondifferencemakers is safe: because these factors play no role in determining whether or not a phenomenon occurs, they can be deformed without affecting a model's predictive power. Or at least, this will hold true provided that we define differencemaking in an appropriate way—a project with which the greater part of this paper will be concerned.

My proposal, summing up the foregoing line of thought, is this: an idealization is safe just in case it distorts exclusively nondifferencemaking elements of the causal process represented by the model. I will call this principle the *differencemaking/idealization thesis*.

A few remarks about the thesis. The first is an objection. If two differencemakers are distorted in compensating ways, the resulting model might, in spite of its getting the differencemakers wrong, be a perfect predictor. Wouldn't idealizing the differencemakers in that case be safe? For that matter, a model might completely misrepresent the differencemakers in the target system yet be a perfect predictor because what it does represent is computationally equivalent, as it were, to the actual system. What to say about that? I say that "getting lucky", though it is indisputably a *route* to success, is not a *strategy* for success of any kind, and so a fortiori, not a strategy for idealization in science. As we are in need of a strategy, we insist on getting the right answers, as modelers say, for the right reasons. Only then are we truly "playing it safe".

Second, as explained in section 2.1, we need a safety criterion that applies not only to explicit idealizations but also to idealizations that are implicit in black boxes. The differencemaking/idealization thesis looks to provide this: it is safe to use a black box that implies a certain causal distortion (such as the absence of certain irrational tendencies in human decision-makers) just in case the distorted factor is not a differencemaker. The black box will then specify behaviors that deviate from actual behaviors only in ways that make no difference to the model's predictions.

Third, in fleshing out the differencemaking/idealization thesis, there is

a choice to make: does it concern differencemaking in the actual causal process, or in the causal process as represented (perhaps incompletely) by the predictor's current knowledge? I will settle ultimately for an epistemically relative characterization.

Fourth, it is important to see that differencemakers and nondifferencemakers are often aspects of the very same properties or things. To take a simple example, it may make a difference to whether or not a window shatters that the brick that strikes it has a certain minimum velocity. But given that the brick is indeed traveling fast enough, its exact velocity will likely make no further difference to the breaking. Here we factor a single causal element—the velocity—into two parts, an approximate value and an exact value within the range allowed by that approximate value, and we discern that one is differencemaking and the other is not. In so doing, we cut across the system's natural ontological joints to see what aspects of the causal process are essential to the shattering. In a slogan: differencemakers are (usually) not things, or even properties of things, but aspects of properties of things.

What, then, is differencemaking? There is more than one notion going under that name in the philosophical literature. None is inherently right or wrong, but one is far better than the others for the purpose of determining when an idealization threatens the accuracy of a predictive model. In what follows, rather than attempting to survey every way of thinking about differencemaking, I will consider two especially prominent approaches: a counterfactually driven notion of differencemaking and a logically driven notion, the latter based on my own theory of differencemaking in explanation. It is the logical version of the differencemaking/idealization thesis, I argue, that provides the correct account of safety.

3. The Counterfactual Approach

3.1 Counterfactual Differencemaking

To characterize counterfactual differencemaking, let us take the *via negativa*. An event can be said *not* to counterfactually depend on a certain aspect of the causal process that brought it about just in case, had that aspect not been present, the event would have occurred all the same. More generally, a kind of event can be said not to counterfactually depend on a certain causal element across a range of initial conditions just in case, for those conditions in which an event of that kind would occur, were the element to be missing the event would still occur, and the same *mutatis mutandis* for initial conditions in which an event of that kind would not occur (that is: were the aspect missing, the event would still not occur).

The counterfactual approach to differencemaking equates the lack of counterfactual dependence with nondifferencemaking. The nondifferencemakers are therefore those whose absence would change nothing with respect to the occurrence of the target phenomena, no matter what the initial conditions. All other causal elements are differencemakers. (Thus for my purposes, an event counts as a differencemaker across a range of conditions even if, intuitively, it makes a difference in only some of those conditions.)

On a counterfactual interpretation of the differencemaking/idealization thesis, then, a causal element can be safely idealized for predictive purposes only if, for every relevant set of initial conditions, were the element not present, the fact of the target phenomenon's occurrence or non-occurrence would remain the same.

That sounds like a good first pass at a recipe for avoiding idealizations that degrade a model's predictive reliability. But it is only a first pass, as there are two serious difficulties with the counterfactual approach. In many cases, it yields unclear verdicts when applied to structural properties, as opposed to discrete events. And in some circumstances, it gives bad advice about

idealization, permitting idealizations that undercut predictive power.

3.2 First Problem: Unclear Verdicts

Let our predictive task be that of answering the question "Does sodium react strongly with water?".[5] This is, of course, a simplified version of the more subtle question "Under what circumstances, if any, does sodium react strongly with water?", which will be answered, as described above, by a model that takes various scenarios—various sets of initial conditions—and predicts either a strong reaction or not for each scenario. (As is typical in setting up such questions, a somewhat artificial criterion will have to be laid down to distinguish "strong" reactions from the rest.)

An exemplary nondifferencemaker for the reactivity question—the kind of thing that can be idealized away without predictive harm—is the detailed structure of the sodium nucleus. Certain facts about the nucleus, such as its stability and its charge, are important, but the rest may be simplified or otherwise ignored by a model whose sole purpose is to predict reactivity. Can the counterfactual approach to differencemaking, channeled through the differencemaking/idealization thesis, replicate this judgment?

To apply the counterfactual approach we must, for any given set of initial conditions, answer the question: Would the sodium sample react strongly with the water if such and such details of the structure of the nucleus were different? Using the possible-worlds semantics for evaluating counterfactual conditionals, that means finding the closest worlds where the nuclear structure is different, and noting whether there is a strong reaction between the sodium and the water in such worlds.

The closest worlds will be those, roughly speaking, in which the nuclear structure is as similar as possible to the actual nuclear structure (though of course different with respect to the aspect whose differencemaking status is

5. I used the same example to explore the deficiencies of a counterfactual approach to differencemaking for explanatory purposes in Strevens (2008), chap. 3.

under consideration), while at the same time having as few ramifications as possible for everything else. A moment of reflection should convince you that it is not easy to say what these worlds are like. Do they have subtly different versions of quantum chromodynamics? But even subtle differences might, if they were significant enough to change the nuclear structure, have dramatic consequences for the behavior of sodium. The sodium atom might not be stable enough to get anywhere near the water. So would sodium be reactive in such worlds? We want to say yes, in order to vindicate the intuitive verdict about the detailed structure's not being a differencemaker for reactivity. Instead, we draw a blank. This suggests that our judgments about differencemaking are not based on the kinds of considerations that determine the truth values of counterfactual conditionals at all.

Here, however, it is important to recall that my goal is to find what I have called a philosophical criterion for safety: a principle that specifies those factors upon which the safety of an idealization ultimately depends. Such a rule need not be one that is applied on a day-to-day basis by working scientists, or indeed, by working philosophers. In the case at hand, modelers need not make their decisions about the safety of idealizing the sodium nucleus by calculating the similarity of possible worlds. They may be using a suite of heuristics that go by some different route, but nevertheless arrive reliably and for systematic reasons at the destination specified by the counterfactual version of the differencemaking/idealization thesis.

The difficulty with the counterfactual version is not that it is hard for ordinary scientists to apply, but that there is reason to think that no matter how expertly it is applied, it will fail to yield the right answer to the question about the nuclear structure's differencemaking status. I say this because it seems quite plausible either that there is no fact of the matter about which worlds are closest, or that in at least some of the closest worlds, chemistry works sufficiently differently that sodium does not react strongly with water. Either way, the counterfactual criterion will answer our question incorrectly:

it will fail to classify the details of nuclear structure as nondifferencemaking.

Let me now turn to an aspect of the sodium atom that clearly does make a difference to sodium's propensity to react with water: its single, loosely bound outer electron. For the counterfactual criterion to render the correct verdict about the electron, it must be the case that in some chemical scenarios at least, were the electron not loosely bound, sodium would not react strongly with water. Again we face great perplexity in determining the relevant closest worlds. Are they worlds in which sodium has a single *tightly* bound outer electron? What laws of nature differ, such that the binding is so strong, and what are the implications for everything else? Or maybe the closest worlds are those where sodium's outer shell has several electrons. But what would bring that eventuality about? Could we identify atoms of this sort as sodium?

There is surely a real prospect—though we are all, in these matters, groping in the dark—that there is no fact as to which worlds are closest. In that case, there is no determinate truth value to the crucial counterfactual, thus no determinate fact about differencemaking. But the loosely bound electron obviously makes a difference to reactivity! The counterfactual criterion is making absurdly heavy weather of what should be a relatively straightforward chemical judgment.

Perhaps the problem lies not with the counterfactual criterion for differencemaking, but with the possible-worlds semantics for counterfactual conditionals? Provided that the conditionals in question are those of ordinary language, I think not. The kinds of questions that the possible-worlds semantics forces on us—what would the world be like if sodium's structure were different in various ways?—are questions that are genuinely relevant to evaluating the ordinary-language conditionals in question. Even if the possible-worlds semantics is not correct, it is quite right in implying that valid counterfactual reasoning about tweaks to the structure of sodium requires the spinning out of an array of alternatives to the actual world's physics. Whatever is the correct story about ordinary counterfactual conditionals, then, will

run into many of the same difficulties as the possible-worlds interpretation, if forced to deliver judgments about the differencemaking status of various aspects of atomic structure.

An alternative is to interpret the differencemaking/idealization thesis using counterfactual conditionals of a more technical sort, such as the model-relative counterfactuals of certain interventionist characterizations of causality (Galles and Pearl 1998). These, it seems, by confining their pronouncements to facts that are implicit in a well-specified model, rather than setting out to explore the space of possible worlds, in large part avoid setting themselves the daunting task of assessing the global implications of counterfactual tweaks. They might indeed provide determinate, correct answers to my questions about differencemaking in the sodium atom; at least, I won't try to argue otherwise. At the end of section 4, however, I will advance a reason to think that even if these conditionals, or any others like them, supply the right answers to questions about the safety of various idealizations, they fail to identify the ultimate determinants of safety. Thus, they fall short of providing a philosophical theory of safety.

3.3 Second Problem: Bad Advice about Idealization

Take some aspect of a causal process that is judged by the counterfactual criterion to be a nondifferencemaker for the target phenomenon. According to the differencemaking/idealization thesis, such a feature can be idealized without compromising a model's predictive reliability. That, however, is faulty advice.

What's guaranteed by the feature's having "passed" the test for nondifferencemaking is that its removal, as carried out in the relevant closest possible worlds, will have no effect on the occurrence of the target phenomenon. What makes a world close is that such a removal is accomplished discreetly—with a minimum of fuss, as it were. That a feature is a nondifferencemaker in the counterfactual sense tells us, then, that certain relatively small changes are

predictively safe, but that is all; the predictive upshot of more radical changes is untested.

Many idealizations, however, are huge departures from actuality, almost to the point of absurdity. In the ideal gas model, we represent infinitely small molecules; in population genetics, infinitely large populations; in microeconomics, ideally rational economic actors. Examining counterfactual removals will tell us little or nothing about the impact of such extreme distortions.

In short: the deliverances of the counterfactual criterion hinge on the consequences of the most conservative transformations, whereas when idealizing we want to know about the consequences of any transformation, no matter how profound. The previous objection showed how challenging it is to distinguish the most conservative transformations from the rest. Making that distinction now turns out to be unnecessary, even counterproductive.

What if the counterfactual criterion is amended to take into account the specifics of the proposed idealization? So rather than asking (for example) "What if the molecules were not that particular size?" we ask "What if the molecules were infinitely small?". That guarantees that the idealization we're envisaging will hold in the possible worlds used to evaluate the conditional. (Never mind that it means we have to propose idealizations before we have distinguished what it is safe to idealize.)

The answer to such a question still won't give us reliable advice on idealizations, however. It will tell us what will happen in the closest possible worlds where the idealization holds, but the causal structure specified by the idealized model might differ in crucial ways from the causal structure of such worlds. It might be, for example, that the most conservative way to idealize a nondifferencemaker is to alter in addition certain differencemaking aspects of the relevant causal process (which changes presumably ameliorate the most radical side effects of the proposed idealization). The closest worlds we are consulting to settle questions about differencemaking will then differ

in crucial ways from our model.[6]

Counterfactual conditionals are good for many things in philosophy. (I've used them myself to define the explanatorily important relation of entanglement.)[7] They may even articulate a philosophically useful notion of differencemaking (although my argument shows that such a notion would be rather limited, going indeterminate for many questions of great scientific interest). But they are not the tool needed to put the differencemaking/idealization thesis to work. Perhaps an alternative criterion for differencemaking can do better.

4. The Logical Approach

The logical approach to differencemaking presented in what follows is derived from the "kairetic" conception of differencemaking that I have developed in other work to diagnose causal-explanatory relevance. Unlike kairetic differencemaking (and indeed counterfactual differencemaking), logical differencemaking is epistemically relativized: what makes a difference to what, in the logical sense, depends on what a scientist knows about a causal process, or more formally, is indexed to an epistemic situation. I will say something further about the connections between the two notions of differencemaking in section 5.

To determine a phenomenon's logical differencemakers for a specified range of scenarios and a given scientist, begin with an accurate representation of everything the scientist knows about the causal production of the phenomenon in those scenarios. This representation provides a basis for

6. You could, of course, nail down everything that matters in the counterfactual antecedent, in effect pinpointing explicitly the closest world you wish to examine. But at that stage, you are evaluating a counterfactual conditional only in the most nominal sense, since you are merely examining the logical consequences of your minutely specified antecedent. What you have is in fact a cumbersome variant of the logical notion of differencemaking to be presented in the next section.

7. In Strevens (2008), Strevens (2012), and Strevens (2014).

predictions that are as good as anyone in the scientist's epistemic situation can reasonably expect to make on the grounds of causal information alone.

Indeed, with one small amendment, the representation can be regarded as itself constituting a predictive causal model for the phenomenon. The amendment is the addition of what I call a causal totality statement, which specifies that there are no further causally relevant facts. In some cases, the totality statement is known to be true; then, of course, it is already present. In other cases, the scientist may strongly suspect that it is false. It must be inserted into the model all the same. It mimics, in effect, the scientist's intention to issue predictions in spite of the incompleteness of their knowledge.

With the totality assumption in place, the representation of the scientist's knowledge acquires the predictive capacity of a causal model: add a set of initial conditions to the representation, and insofar as you can derive anything determinate about the occurrence of the target phenomenon, you have a prediction—the best prediction, as I have said, that the scientist could in principle make on the grounds of what they currently know. I call this the *epistemically maximal model* for the scientist in question. By assumption, the scientist is ready to build a useful predictive model (or the question of safety simply doesn't arise), so I take it that the epistemically maximal model has considerable predictive power.

Although the epistemically maximal model is derived from the scientist's epistemic state, it is not their model. It is a philosopher's construction, a part of a philosophical theory of what makes a predictive idealization safe. The scientist has yet to build their model. Their access to the maximal model may in fact be rather limited. They might not "know what they know"—for example, they might be highly confident about the existence of some causal feature that is not in fact present. They take this to be a part of their knowledge, but it is not. Or even if they have a good grasp of what constitutes their knowledge, they might lack the means to extract a prediction, if that knowledge is complex. The problem of determining what their knowledge

implies might be, for them, intractable.

The point of starting with the epistemically maximal model is not to set out a method for the scientist to follow but to articulate what I have called a philosophical criterion for determining safety—that is, a principle that pinpoints what the scientist, given the epistemic foundations upon which they hope to build their causal model, can safely idealize. It may not be easy for them to learn these facts about safety, and as I have remarked above, when they succeed it will typically be by using some heuristic, not by the first-principles approach that invokes the notion of logical differencemaking defined in this section. But the value of their efforts must ultimately be judged against the definition.

On with the characterization of logical differencemaking, then. Having assembled the epistemically maximal model, the next step is to see what can be deleted from the model without affecting its predictions, that is, without changing the answers it gives about the occurrence of the target phenomenon, "yes" or "no", for any of the sets of initial conditions within the model's intended range of use. "Deletion" here means literally that: simply removing information from the model, so that it says strictly less about the causal process in question than it did before the removal. A deletion might omit mention of some causal factor altogether; the model then becomes agnostic as to the presence of that factor.[8] Or it might make a description of a property less specific—replacing an exact velocity for a hurled brick, for example, with a range of velocities spanning the exact value. The model now says less about the brick's velocity than it did before.

Even more ambitiously, a deletion might replace a model's specification of the states of a large number of individual entities with a statistical profile of the same states. Done right, this "statisticalization" leaves you with strictly less information about the causal process in question. The removal procedure,

8. Unless the factor is entailed by other elements of the model, as might be the case if it is an intermediate stage in a process whose causal precursors remain spelled out in the model.

as I have described it, is essentially one of abstraction. A causal model is made more abstract by replacing some sentence or equation in the model, or some set of sentences or equations, with other representations that (a) are entailed by the originals, and (b) concern the same subject matter, or a subset thereof.

The idea behind logical differencemaking, roughly, is this: if the description of some aspect of the causal process can be removed from the model without affecting any of the model's predictions over the relevant range of initial conditions, that aspect is a nondifferencemaker for the target phenomenon. Otherwise, it is a differencemaker. Two further conditions must be added, however, before this characterization is ready for action.

First, the deletion must not be so severe that what's left is no longer a causal model. You could maximize information removal by replacing everything in the model with a lookup table relating initial conditions to the yes or no answers given by the original model, without predictive loss. Such a table might be useful in certain circumstances, but it is not a representation of the causal process producing the target phenomenon, hence not a causal model (or indeed, any kind of model).

Second, and more subtly, the model, like any predictive causal model, must have some unity or cohesion. The internal workings of the model could in principle be replaced with a binary disjunction, in which the original workings constitute one disjunct and some other causal process with the same consequences constitutes the other. But that would yield what is intuitively a disjunction of causal models, describing two different kinds of processes, not a single model.

The rationale for these qualifications lies outside the scope of this paper, but it is not too difficult to discern: we want to build on our causal knowledge. If we were finished with scientific inquiry and cared only about prediction, the lookup table would suffice for all our needs, as would an appropriately formulated disjunction of models. In our usual, limited epistemic situation, we value discrete causal representations because we imagine that we will

amend, improve, adapt, and analogize them in the service of further scientific achievements. The same standards should be applied to the maximal model.

To determine logical differencemaking, then, abstract the epistemically maximal model as far as possible without violating the requirements enumerated above: that the model continue to yield the same predictions and that it continue to constitute a cohesive causal model. At that point, everything that has been removed is declared to be a logical nondifferencemaker; only logical differencemakers remain.[9]

Logical differencemaking can equally well be defined without using the procedural idiom: a factor is a logical non-differencemaker just in case there exists some abstraction of the epistemically maximal model that is itself a cohesive causal model of identical predictive portent and in which that factor does not appear. This static formulation is less vivid, but it makes it quite clear that the facts about logical differencemaking relative to a body of knowledge are fixed by the logical consequence relation and the criterion for being a cohesive causal model—and thus that they do not depend on the counterfactual labors of some epistemically maximal modeler.

To complete the story, I invoke the differencemaking/idealization thesis: the aspects of causal structure that our scientist may safely idealize are the logical nondifferencemakers. In other words, the accurate information that has been removed from the model can safely be replaced by intentional misrepresentations. The only constraint is that these misrepresentations, the idealizations, must be logically consistent with the differencemakers.[10] If we

9. For technical reasons related to note 8, a more careful presentation of the logical approach to differencemaking will represent causal structure not using a single causal model but using a set of such models representing different but overlapping segments of the entire known causal structure. A factor is a nondifferencemaker in the logical sense if it can be removed from all such models without changing the model's predictions (Strevens 2008). I suppress this additional layer of complexity in the main text.

10. The reason that such idealizations are safe, in the most schematic terms, is this. The abstracted model, containing only logical differencemakers, supplies all the information in the scientist's knowledge base relevant to the predictive task. Adding anything further to the model cannot usefully supplement the information, then. But nor can it subtract from

were able, in the process of abstracting a model of a gas, to replace an exact value for molecular size with a range of values, then we can idealize by specifying any other value for molecular size that lies within that range—perhaps zero. If we were able, in the process of abstracting a model of a population of organisms, to remove any information as to whether generations overlapped or not, then we can idealize by specifying non-overlapping generations. And so on.

The logical differencemakers, by contrast—the elements of the causal structure that remain after the abstraction procedure described above—must be faithfully represented in a predictive model, if it is to fully realize the predictive potential of the scientist's causal knowledge.

Let me return to the question of sodium's reactivity, to see how the logical approach to differencemaking handles the case. We begin with a model of reactivity that presumably predicts a strong reaction with water. We ask: what can be expunged? Subtract details of the sodium atom's nuclear structure, and the model continues to predict reactivity, provided that you leave behind enough information to entail the nucleus's charge and the stability of the atom.[11] Subtract information about the outer electron, however—either its existence or the weakness of the bond holding it to the nucleus—and your model will no longer say much that is determinate about sodium's reaction with water. Thus, the details of nuclear structure are nondifferencemakers while the facts about the outer electron are differencemakers. Idealize the former—representing the nucleus as a point charge, say—but leave the latter alone, and you will have a simplified model of sodium that is still a good predictor for the purpose at hand.

---

its predictive usefulness. That is because the abstracted model supplies the information by deductively entailing it, and deductive logic is monotonic. (When the model is stochastic, what is entailed is a probability, which is then interpreted as a yes or no answer.) Adding something to the model, even something false, therefore cannot undo the entailment.

11. Take away enough information, and the representation of the nucleus becomes something approaching a black box. It is up to the scientist to decide whether a black-boxed model is sufficiently cohesive and unified for their predictive purposes.

Characteristically, we do not make these judgments by scrupulously following the recipe laid out above, starting with an epistemically maximal model for sodium's behavior, and so forth. The maximal model, which for us today would include all of the relevant quantum chemistry, is too complex to easily manipulate (indeed, it is very difficult to manipulate at all). But we have our heuristics. We can see from the overall form of the dependences involved that nuclear structure won't matter much, and that the outer electron is doing a lot of work, even if we can't carry out the calculations that would provide a rigorous proof. Our heuristics show us what will and won't be a logical differencemaker, without our having to crank through the definition of logical differencemaking.

That logical differencemaking is relative to a knowledge situation, and invokes a causal totality statement that might well be false, raises the following concern: a causal factor might count as a nondifferencemaker in the context of a scientist's current knowledge, but as a differencemaker once more is known. Suppose, for example, that the scientist knows of a mechanism by which a factor X has only a slight effect on the relevant course of events, not serious enough to affect the epistemically maximal model's predictions. X will qualify as a nondifferencemaker for the predictive task. But there might be some other, unknown mechanism by way of which X has a considerable impact on the occurrence of the target phenomenon. Were this mechanism known, X would certainly qualify as a differencemaker. Yet the differencemaking/idealization thesis, interpreted as I have proposed here, will tell us that the predictive model can safely be idealized with respect to X—that, for example, the scientist can safely assume, in the predictive model, that X is absent or altogether causally unconnected to the target phenomenon.

Is that good advice? In fact, it is. Given what the scientist knows, and is therefore capable of including in their predictive model, X can indeed be distorted or ignored. For now, it is safe to idealize X. When more becomes known, it will become unsafe to idealize X. The logical approach delivers

these verdicts, and they are, even if disconcerting, correct.

For various reasons, a scientist might choose, when building a predictive model, to ignore some of their knowledge of the underlying causal structure. Including every known facet of the structure might, for example, simply overwhelm the scientist's computational capacities; to get any prediction at all, then, they must pick and choose what they take to be adequate predictors, knowing that their model will fall short in certain ways.

What may be safely idealized in such a model? The logical difference-making approach supplies an answer. To assess safety, begin not with all the scientist's causal knowledge, but with that subset of their knowledge that they propose to use to build their model. Add a causal totality statement (known, in this case, to be false). Apply the abstraction procedure to determine differencemakers and nondifferencemakers relative to the knowledge subset. It is safe to idealize any of the nondifferencemakers.

Of course it may be that some of these nondifferencemakers qualify as differencemakers relative to the scientist's total knowledge (in the same way that some nondifferencemakers relative to their total knowledge will be differencemakers relative to more inclusive sets of information still). In following my advice, the scientist may therefore be idealizing a factor that they know to be a differencemaker (or to put it more carefully, a factor that is, from the perspective of their entire knowledge, a differencemaker). Such factors may nevertheless be idealized safely, for the same reason given in the reply to the objection above: because of the limitations on what goes into the predictive model, there is nothing in the model by way of which the idealization can affect the predictions in question.

A rather different way for a scientist to ignore some of their causal knowledge is to black-box certain parts of the target system. Strictly speaking, a black box does not represent any aspect of underlying causal structure and so cannot distort that structure. As I noted in section 2.1, however, some black boxes are widely understood to imply distortions; a black box that replicates

the decision-making of the perfectly rational *Homo economicus*, for example, is often taken to idealize human decision-making by supposing, falsely, that humans are perfectly rational.

In section 2.4, I suggested that such idealizations are safe just in case the distorted factors are not differencemakers. Let me make good on this proposal by showing how the logical approach to differencemaking handles black boxes.

Suppose that our scientist has declared their intent to black-box certain subsystems or processes; they may know something about the causal structure of these subsystems, but they will omit what they know from their predictive model, substituting a bare function relating inputs to outputs—the black box itself.

To provide a safety criterion for such a model, amend the logical criterion for differencemaking as follows. Begin, as before, with an epistemically maximal model, representing all the scientist's relevant knowledge. Then black-box the model in accordance with the scientist's intentions, replacing any knowledge of the structures to be black-boxed with a bare input/output function.

The usual abstraction procedure can now be applied to this black-boxed model. (The details are discussed further in Strevens (2016).) Suppose, by way of illustration, that a black box in the model is "deterministic", supplying a single determinate output for every input. The precise value of the output may not matter for predictive purposes; it may be enough to know an approximate value. In the course of abstraction, the black box will be switched out for a more abstract black box, saying strictly less about what output results from any given input. When abstraction is done—when all possible such substitutions have been made—you have in place of your black boxes what might be regarded as black box templates. Any way of filling out those templates will yield the same predictions. The templates may safely be replaced, then, with determinate input/output functions that fit the template but are known to be

inaccurate—black boxes whose description of the relation between inputs and outputs is at best only approximately correct.

Some such replacements, like the black box that represents the consequences of perfectly rational decision-making, may imply distortions of causal structure. These black-box idealizations are the safe ones.

Finally, I can make good on my claim in the previous section that there is something objectionable about combining the differencemaking/idealization thesis with a counterfactual criterion for differencemaking, no matter what flavor of counterfactual conditional is employed. By a counterfactual approach to differencemaking, I mean one that, in order to determine whether a causal factor makes a difference, examines—by way of possible worlds, interventionist causal models, or some other structure—the consequences of possible but non-actual alternatives to that factor. In the case of sodium, for example, a counterfactual test of any stripe will ask us to flesh out a coherent story in which the sodium atom has one or more counterfactual structures, or operates according to one or more counterfactual principles.

The effectiveness of the logical approach to differencemaking as a test for safety shows not only that the consideration of such alternative structures and principles is unnecessary to determining safety, but that safety does not in the end depend at all on the behavior of these alternatives. It depends only on the question of which parts of the epistemically maximal model are doing the work of entailing the information used to make predictions. Even if a counterfactual test delivers the right advice about safety, then, were it to be considered as a philosophical criterion for safety—a principle that spells out what ultimately makes an idealization safe or unsafe—it would falsely imply the importance of things that are in fact explanatorily irrelevant to safety. If any such tests exist, they can be regarded only as useful heuristics.

5. Idealization and Differencemaking in Prediction and Explanation

Scientific explanation, most contemporary philosophers agree, makes use of a notion of causal relevance, choosing certain causal factors—conditions, structures, regularities, and laws—as explainers of a phenomenon because they are causally relevant to that phenomenon in a certain way. Prediction using causal models also invokes a standard of causal relevance, I have argued in this paper, to discriminate between causal factors that can and cannot be safely idealized.

The relevant factors in explanation are those that really matter for explanatory purposes. The relevant factors in prediction are those that really matter for predictive purposes. There is no obvious reason why these two forms of relevance should not come apart. There are many coherent philosophical notions of causal relevance in the literature, some built around the idea of differencemaking and some not, appealing to statistics, counterfactuals, nomological relevance, or—like mine—simple logic. It is surely to be expected that the notion required by explainers will differ in some respects from the notion required by predictors.

But in fact, the approach to relevance that works for prediction is strikingly similar to the account of differencemaking that determines—if my work on the topic is correct—explanatory relevance.

Hempel (1965, §2.4) has frequently been criticized for asserting a symmetry between prediction and explanation. He was indeed wrong to leave causation out of the picture. The barometer's needle drop does not explain the storm, and the length of the flagpole's shadow does not explain the pole's height, however predictive they may be. But a certain, carefully circumscribed version of Hempel's thesis looks like it might be correct: the factors that matter for explanation are precisely those that matter for *causal* prediction. That is, an aspect of the causal process producing a phenomenon plays a part in explaining that phenomenon just in case it plays an essential part in predicting that phenomenon *on causal grounds*.

The logical approach developed in this paper is not, on close examination, strictly identical to the kairetic criterion for explanatory relevance. Let me examine the two principal differences between the predictive and explanatory criteria, showing that they are philosophically insignificant; I will thereby preserve the unexpected Hempelian moral of my story.

The first difference is that differencemaking for the purposes of prediction is relative to the criterion for deciding whether a predictive model says "yes" or "no". Suppose that we are working with a model that, for any set of initial conditions, ascribes a precise probability to the target phenomenon. A standard yes/no criterion would deliver a "yes" just in case that probability is equal to or greater than one-half. A factor that increases the probability only when it is already over one-half, then, is not a predictive differencemaker: with or without that factor, the model will deliver the same yes/no answers. But on many accounts of probabilistic explanation, including my own, it is an explanatory differencemaker.

This divergence between prediction and explanation is rather superficial, however: it is an artifact of the demand for binary answers, rather than a feature inherent to the art and craft of prediction. In most real-world contexts, a predictive modeler will take account of the magnitude of the probability ascribed to the target phenomenon; they will take account, then, of probability-raisers (and they will be reluctant to idealize them away).

The second difference between predictive and explanatory differencemaking is that the predictive sort but not the explanatory sort is relative to a modeler's epistemic situation. In the explanatory case, a nondifferencemaker is something that can be removed from a model representing the complete truth about the causal production of the target phenomenon without changing the probability ascribed to the target phenomenon.[12] In the predictive case, a nondifferencemaker is something that can be removed from a model that

12. This and subsequent formulations of differencemaking are simplifications adopted for clarity's sake, along the lines remarked in note 9.

summarizes the predictor's current causal knowledge without changing that probability. As I observed earlier, some nondifferencemakers in the predictive sense will be differencemakers in the explanatory sense, namely, those that affect the probability by way of a causal pathway that is as yet unknown.

Again, though, this difference between predictive and explanatory differencemaking is not as deep as it may first appear. There is nothing to distinguish explanatory differencemaking and predictive differencemaking for a predictor who knows everything about the structure of the causal system whose behavior they seek to predict or explain. To a causally fully informed scientist, in other words, explanatory and predictive causal differencemaking, and therefore relevance, are identical. That is not Hempel's symmetry thesis, but it has much of its character. The grand old philosopher of science was onto something after all.

To conclude, a few words on the purpose of idealization. I have already sketched one such function: to simplify models in order to make them easier to use. Potochnik (2017) has suggested that idealizations can assist in the efficient representation of causal patterns; Bokulich (2016) says the same about the representation of patterns of counterfactual dependence, among other things. Finally, in my own work on explanatory idealization (Strevens 2008, chap. 8), I have proposed that ostentatiously fictitious idealizations communicate forcefully to a model's consumers—other scientists—that certain causal elements that might well have been thought to play a decisive role in determining the occurrence of the target phenomenon are in fact nondifferencemakers.

These writers are primarily concerned with explanation, but the very same functions may be performed, in each case, by predictive idealizations. More generally, I suggest that idealization will play parallel roles in the predictive and the explanatory enterprises. There may be some divergence in idealizing taste—a predictive modeler may be more interested in idealizations that make a model easier to use, while an explanatory modeler may be more interested

in idealizations that make a model easier to comprehend, insofar as those two qualities come apart—but presumably this will be at most a difference in degree.

In any case, no matter what the function or functions of idealization, the question of safety will arise. (An unsafe explanatory idealization weakens an explanation by removing explanatorily relevant information.) And the criterion for safety to which an idealizer must ultimately conform, whether they are explaining or predicting, will turn on the the facts about differencemaking—with differencemaking interpreted not by way of a counterfactual criterion but in accordance with the logical approach.

References

Appiah, K. A. (2017). *As If: Idealization and Ideals*. Harvard University Press, Cambridge, MA.

Bokulich, A. (2016). Fiction as a vehicle for truth: Moving beyond the ontic conception. *The Monist* 99:260–279.

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

Elgin, C. (2017). *True Enough*. MIT Press, Cambridge, MA.

Galles, D. and J. Pearl. (1998). An axiomatic characterization of causal counterfactuals. *Foundations of Science* 3:151–182.

Hempel, C. G. (1965). Aspects of scientific explanation. In *Aspects of Scientific Explanation*, chap. 12, pp. 331–496. Free Press, New York.

McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science* 16:247–273.

Mäki, U. (1992). On the method of isolation in economics. In C. Dilworth (ed.), *Idealization IV: Intelligibility in Science*, volume 26 of *Poznań Studies in the Philosophy of the Sciences and the Humanities*, pp. 317–351. Rodopi, Amsterdam.

Nowak, L. (1992). The idealizational approach to science: A survey. In J. Brzezinski and L. Nowak (eds.), *Idealization III: Approximation and Truth*, volume 25 of *Poznań Studies in the Philosophy of the Sciences and the Humanities*, pp. 9–63. Rodopi, Amsterdam and Atlanta, GA.

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

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, MA.

———. (2012). The explanatory role of irreducible properties. *Noûs* 46:754–780.

———. (2014). High-level exceptions explained. *Erkenntnis* 79:1819–1832.

———. (2016). Special-science autonomy and the division of labor. In M. Couch and J. Pfeifer (eds.), *The Philosophy of Philip Kitcher*. Oxford University Press, Oxford.

———. (2017). How idealizations provide understanding. In S. R. Grimm, C. Baumberger, and S. Ammon (eds.), *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*. Routledge, New York.

———. (2019). The structure of asymptotic idealization. *Synthese* 196:1713–1731.

Suárez, M. (1999). The role of models in the application of scientific theories: Epistemological implications. In M. S. Morgan and M. Morrison (eds.), *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press, Cambridge.

Weisberg, M. (2007). Three kinds of idealization. *Journal of Philosophy* 104:639–659.

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