Synthese (2014) 191:2673–2693
DOI 10.1007/s11229-014-0408-3

Model change and reliability in scientific inference

Erich Kummerfeld · David Danks

Received: 29 January 2013 / Accepted: 27 January 2014 / Published online: 26 February 2014

Abstract One persistent challenge in scientific practice is that the structure of the world can be unstable: changes in the broader context can alter which model of a phenomenon is preferred, all without any overt signal. Scientific discovery becomes much harder when we have a moving target, and the resulting incorrect understandings of relationships in the world can have significant real-world and practical consequences. In this paper, we argue that it is common (in certain sciences) to have changes of context that lead to changes in the relationships under study, but that standard normative accounts of scientific inquiry have assumed away this problem. At the same time, we show that inference and discovery methods can "protect" themselves in various ways against this possibility by using methods with the novel methodological virtue of "diligence." Unfortunately, this desirable virtue provably is incompatible with other desirable methodological virtues that are central to reliable inquiry. No scientific method can provide every virtue that we might want.

1 Introduction

Essentially all normative accounts of science, as well as the practices of many scientists themselves, reflect a common assumption: namely, that the correct models of the world are stable through time. That is, there is frequently an assumption that the *relationships* between elements in the model are stable over time, though the values of those elements can of course change. For example, one might believe or assume that the (probabilistic) causal relation between a light switch and the state of the lights is

E. Kummerfeld (✉) · D. Danks
Carnegie Mellon University, Pittsburgh, PA 15213, USA
e-mail: ekummerfeld@gmail.com; ekummerf@andrew.cmu.edu

D. Danks
e-mail: ddanks@andrew.cmu.edu

stable, even though the particular states of the switch and light obviously change over time. This assumption manifests itself in many ways in scientific practice, including the reuse of data from earlier experiments, the drive to subsume earlier theories in later ones, and the rejection of theories that failed earlier experimental tests.

For many sciences and many real-world contexts, however, this assumption is false. More generally, our understanding of the world—causal, social, physical, and other—is through models that ignore (necessarily, we argue below) broader contextual factors that are potentially relevant, and so when those factors outside of the model change, the correct model of the world can also change. We will refer to the former phenomenon—change in the unmodeled factors or conditions—as *context change*, and the resulting change in which model is correct (when that occurs) as *context-driven model change*. We here focus on this problem from the perspective of scientific inquiry and methodology; we are less interested in the models themselves (though much of the discussion below uses the language of models), and more in the methods that scientists should use to respond to the possibility of this context-driven model change. When there is a change, for example, in the causal relation between the switch and lights (e.g., the power goes out and thereby breaks the causal connection), we require methods that detect and respond to (or otherwise are robust against) such changes.

Such changes pose a significant challenge to many standard scientific practices. For example, if one study concludes that C causes E and another study concludes that C does *not* cause E, then typical scientific practice is to conclude that one (or both) of the studies suffered from some sort of flaw, either in study design or data analysis. If, however, the correct model can vary between contexts (e.g., the correct model is different at different times), then such a conclusion is unwarranted: it is quite possible for *both* studies to be correct at the times each was run.[1] Different methods for scientific inquiry are required if we allow for the possibility of context-driven model change. We provide a schema for principled methods that can solve this problem (Sect. 5), but protection against errors due to context-driven model change comes at a cost. These methods exhibit a novel methodological virtue that we call *diligence*—roughly, the errors made in inquiry or by the learning method are bounded when model change occurs. Unfortunately, diligence is inconsistent with one of the most commonly desired methodological virtues, consistency; there is an irreconcilable tension, and so necessary trade-off, between these methodological virtues (Sect. 5.1).

Before moving to this positive project, however, we first need to show how context-driven model changes emerge and the errors that they can introduce in scientific understanding and practical applications. We provide a more precise specification of the problem of context-driven model change (Sect. 2), and then show that this problem has real, practical impact. Context change is not rare or insignificant, but rather very real: there are multiple real-world examples of correct models changing in significant ways (Sect. 3), and standard normative accounts of scientific methodology neglect this possibility (Sect. 4).

[1] Similarly, if contexts can change, then replications of experiments may have different evidential value than is typically thought.

2 Context-dependence of models

Our principal interest is in models of the world, where we use the term 'model' in a general way to refer to anything that supports prediction, explanation, and control. We deliberately use 'model' in this broad way in order to be agnostic about the exact nature of our scientific and everyday models; our arguments are consistent with any understanding of models in which they provide structured understandings of the world that support key functions. Models necessarily include only a subset of the possible factors or variables, and so inevitably involve some abstraction from the precise, messy details of the system or situation under study (Cartwright 1983, 2007). For both practical and theoretical reasons, we simply cannot include everything in a model, but rather must omit some variables, details, or levels of complication. We will refer to these omissions as the *context* of the model—the elements of any particular situation that are exogenous to the model. Elements of the context might matter for the system under study, but they need not.

One aspect of learning about the world involves finding the best model from a collection of mutually exclusive models, all of which have the same context C. That is, we are often trying to decide which of several models is the correct one. We will refer to this task as finding the *target model* in a *framework*. Roughly, a framework is a set of possible models for some situation and context, and the target model is the best (or one of the equally-best) model according to some defensible measure of a model's quality (e.g., truth, accuracy, etc.). As a simple example, the target model for the lights in one's office is (in everyday contexts) presumably $Switch \rightarrow Lights$, where the causally relevant factor $Power$ is in the context. We are deliberately and explicitly agnostic in this paper about the proper measure of a model's quality: in particular, we take no stand about whether methods of scientific inquiry lead to "true" (in some sense) models, as opposed to ones that are pragmatically useful, defeasibly justified, empirically adequate, or have some other desirable property. As such, the issues that we discuss here are largely orthogonal to those usually studied under the heading of (epistemic or scientific) "contextualism."[2] Whether one is, for example, a fallibilist about knowledge or models in any particular context does not necessarily imply anything about the possibility of context-driven model change (over time), or algorithms that could be used to detect it, or alternative scientific inference methods that are robust to its possibility.

We are principally interested here in what we will call *context-driven model change*: changes in the context that result in a change in the relationships captured in the target model.[3] For example, the target model for one's office changes from $Switch \rightarrow$

[2] Of course, there is agreement with those positions that models are context-relative, but even that agreement is tempered by the fact that we focus on different aspects of context. In particular, we are interested in contextual factors that can change the relationships under study, rather than ones such as the pragmatic desires or goals of the scientists. We suspect that most (epistemic or scientific) contextualists would be quite amenable to the conclusions that we reach in this paper, but we think that our primary focus is quite different from their usual concerns.

[3] Although it will not play a role in this paper, we should note that context-driven model change is framework-relative for two distinct reasons. First, context-driven model change requires that the target model change is due to a change in the context. Since every model in a framework has the same context

Light to *Switch Lights* when the context changes due to the power going out. Other everyday examples of context-driven model change are easy to find, such as a calculator whose '+' key breaks, or a laptop whose battery no longer holds a charge. In all of these cases, there is a stable model M_1 at time t_1, and a different model M_2 at another time t_2. This change can occur because of a single event (e.g., the key breaking or the power going out) or because of a gradual change in the underlying system (e.g., the computer battery slowly losing its ability to retain a charge).

Crucially, context-driven model change includes only cases in which the between-element relationships change; changes over time in the values of model elements (e.g., a light switch changing from *up* to *down* and back again without affecting the causal relationship between *Switch* and *Light*) are not, as we understand it, context-driven model change. More practically, we are not interested in methods for rapid or complex inference *within* a given model; our concern is not about the difficulty of inference in non-linear models or the surprises that can arise through non-monotonic reasoning. Rather, our concern is about the possibility that the relationships in the target model—whatever they might be, and whatever methods we use to do inference about them—can change due to changes in the context (i.e., the relevant factors that are not included in the model).

One response to this possibility is to try to expand the scope of the model to include the relevant aspects of the context. But we contend that all models necessarily have a nonempty context, as noted by numerous authors (Cartwright 1983, 2007): "A 'model,' in the common use of the word, is an idealized representation of reality that highlights some aspects and ignores others" (Pearl 2000, p. 202). Some factors must always be left out or regarded in an idealized manner. For some model M with context C, we can expand it to model M^* by moving some of the factors in C into M^*, but we cannot include *everything*; M^* will necessarily still have a non-empty context. For example, even if we include *Power* in the *Switch & Lights* model, there will still be other unmodeled factors (e.g., whether the bulb filament is intact, or whether the wire from the switch to the bulb is broken). This context-dependence is the inevitable product of the complexity of the world and our boundedness along many dimensions. We can never have a "complete" model, just as no map can represent every aspect of reality; every model has a non-empty context. Arguably *anything* omitted from the model (assuming it is in the appropriate physical light cone, of course) could end up being a necessary part of context-driven model change, given sufficient freedom in changing the context.

In this paper, we are focused on scientific inquiry, and in particular on methods for the discovery of scientific models. Some sciences have started to come to grips with the challenge of context-driven model change (Sects. 3.1, 3.2), but there is little understanding of the general scope and form of the problem. At a high level, all

Footnote 3 continued
(e.g., every model has the same set of variables), the same changes "in the world" can produce context-driven model change relative to one framework but not relative to another, depending on whether the changing factors are in the context or in the framework's models. Second, since the target model is the "best of the bunch," whether the target model changes can depend on the competition (i.e., the other models in the framework).

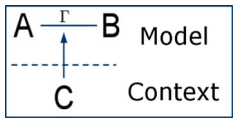

Fig. 1 Schematic structure for context-driven model change

context-driven model change arises when a change in a contextual factor produces a change in the modeled relationships. For example, a change in the contextual factor of the building power (by the power going out) produces a change in the modeled causal relationship between $Switch$ and $Lights$: the target causal model changes from $Switch \rightarrow Lights$ to $Switch\ Lights$.[4]

We can abstractly represent context-driven model change using Fig. 1. Suppose our framework is the relevant models over $\{A, B\}$, where C is some factor in the context.[5] Suppose further that C influences (whether causally, definitionally, logically, or in some other way) property Γ of the A, B relationship (e.g., its existence, informational content, causal direction or strength, etc.). Given some particular target model (i.e., some specification of Γ), changes in A's value can lead to changes in B's value without having any *model* change. However, if C changes between times t_1 to t_2, then the Γ-property can change, and so the target model will change.

Context-driven model change can occur even when we (as scientists) have made no errors in our scientific inference; it can happen simply because our models must exclude some factors, and those factors could end up mattering. Of course, context-driven model change can *also* occur when we make certain kinds of errors in our scientific reasoning. An interesting "special case" is when our framework (i.e., the set of possible models that we entertain) mistakenly conflates heterogeneous (with respect to the modeled relationships) sub-types. Since the sub-types are distinct, the correct target model can easily depend on the distribution of sub-types in the population. But the conflation of those sub-types implies that information about the distribution is contextual, and so changes in that distribution are context changes that can produce model change.

For example, suppose one included in the context (rather than the model) the distribution of whether particular *Staphylococcus aureus* bacteria have the genetic structure to be antibiotic-resistant. In that case, populations of staph—whether regular or antibiotic-resistant—will be picked out by a single variable $Staph$, or more precisely, a variable denoting the population size (corresponding to B in Fig. 1). Because individuals who are different (i.e., have different target models) are grouped together, the overall target model can change as the proportions of those individuals change. For example, $Penicillin$ (A in Fig. 1) is a significant inhibitor of nonresistant staph, but has essentially no effect on resistant strains. For years, the target causal model was $Penicillin \rightarrow Staph$, as it was in fact the case that penicillin treatments led to reductions in the amount of staph in an individual. Once penicillin became mass-produced,

[4] This example makes the framework-relativity of context-driven model change quite clear, as there would be no change in the target model if our framework included the causal model $Switch \rightarrow Lights \leftarrow Power$.

[5] For example, the framework might be the set of causal models: $\{Switch \rightarrow Lights, Switch\ Lights, Switch \leftarrow Lights\}$, and C might be the $Power$ state.

however, resistant strains of staph became more common and so the inhibitory strength of *Penicillin* in the target causal model became significantly lower, perhaps approaching zero (Davies and Davies 2010). That is, the proportion of antibiotic-resistant bacteria (C in Fig. 1) changed, and so the presence of a causal edge (Γ) changed as well.

There has been some previous work in philosophy of science on the types of context change and context-driven model change that concern us here. Edmonds (2007) explicitly discusses the problem of modeling context-dependent causal processes, though those analyses are relatively informal and focus on a different understanding of 'context.' There has also been some formal work on the context-dependency of causal models, but that has been more narrow in scope than our approach. Most notably, Glymour (2011, 2008) identified the important role of interactive causation in biological models, and argued that we should increase model complexity to account for these complex interactions, rather than leaving them in the context of our models. We discuss this approach—what we call the "model expansion" approach—in Sect. 5, but argue that it does not solve all of the scientific inference challenges. Moreover, there are other possible responses to context-driven model change.

3 Context-driven model change in scientific practice

Context-driven model change is clearly a theoretical possibility, but that does not imply that it is necessarily a challenge for actual practice. Unfortunately, context-driven model change is neither so rare, nor so small in magnitude or impact, nor so easily detectable as to make its relevance to scientific practice moot. In fact, context-driven model change is arguably inevitable in some scientific domains, as it requires only that there be factors that are exogenous to the scientists' models but nonetheless have an important influence on the modeled relationships. Such omissions can be unavoidable in practice, simply because of the coarseness of the objects and features studied in that domain. In particular, biology, the atmospheric sciences, and economics are good examples of fields that arguably cannot solve (in an *a priori* manner) the problem of heterogeneous populations (Eells 1991): biologists must conflate populations of organisms with non-identical DNA; atmospheric scientists must aggregate large volumes of space; and economists must aggregate independent economic agents into economies. Context change events are thus almost certainly inevitable for these sciences, and so it is unsurprising that two of them have developed domain-specific methods to try to accommodate the possibility of context-driven model change.

3.1 Biology

Rapid evolution in organisms, regime changes in population dynamics and ecosystems, and norms of reaction are some of the most prevalent examples of context-driven model change in biology. In general, biologists respond by expanding their models: they identify the relevant contextual (i.e., exogenous) factors that result in changing relationships, and then incorporate them into their models.

Rapid evolution with respect to the time-scale of a causal model can change the causal relationships and interaction patterns in the model, such as in the evolution of antibiotic-resistant bacteria (Davies and Davies 2010) and evolution in organisms due to trophic interactions (Duffy and Sivars-Becker 2007; Yoshida et al. 2003, 2007). If we do not include the possibility of rapid evolution in our models, then those causal models (of the relationship between, e.g., bacteria populations and antibacterial agents, or two distinct populations in a predator-prey or host-parasite relationship) can exhibit context-driven model change. That is, we can fail to track the causal relationships in the world because important aspects have been relegated to the context. Biologists have largely adopted the domain-specific response of attempting to identify situations in which rapid evolution may play an important role and then explicitly modeling the rapid evolution in such situations (i.e., expanding their models).

Norms of reaction (Cooper and Zubek 1958; Scheiner 1993; Sarkar and Fuller 2003) have yielded a similar response. A norm of reaction is a function that describes how, for a particular genotype, variations in the value of certain environmental factors (such as average temperature) result in variations of certain phenotypes (such as height). Early biological models of populations of organisms and their environments did not incorporate the influence of the environment on the $Genotype \rightarrow Phenotype$ relationship, and so those models exhibited context-driven model change. Today, norms of reaction are a recognized complication for modeling in biology, both to discover them in particular cases and to use them appropriately in biological models.

Finally, regime change in models of ecosystems and population dynamics prompts both model expansion and novel detection methods as responses to context-driven model change (Scheffer and Carpenter 2003). Regime change occurs when an ecosystem switches from one relatively stable profile of organism populations to another. As before, scientists are sometimes able to identify the relevant exogenous factors so that they can be included in their models (Estes 2011; Scheffer and Carpenter 2003). In other cases, the potentially relevant exogenous factors for regime change are unknown, or there are too many to feasibly survey them. Thus, there has also been significant work in early detection and prediction of context-driven model change (Carpenter et al. 2011); these methods are more domain-general than incorporating exogenous contextual factors since they depend partly on changes in observed statistics, but they remain focused on biological models. We show how to generalize this idea in a completely domain-general manner in Sect. 5.

3.1.1 The Canadian cod fishery collapse

A striking real-world example of the dangers and difficulties of regime change (i.e., context-driven model change) is the 1992 North Atlantic cod (*Gadus morhua*) fishery collapse that devastated the coastal economies of Newfoundland and Labrador (McGuire 1997; Kurlansky 1997; Finlayson 1994).[6] Because fishery scientists

[6] Of course, there are many different issues that the cod fishery collapse illuminates, such as the fact that even central planning and control can be insufficient to prevent a tragedy of the commons. We focus here on the model change aspect, but also think that this is a rich case study that has been insufficiently studied in philosophy of science.

assumed that there was no context-driven model change, the cod population is now less than 1% of its original size, and the ecosystem itself has changed so dramatically that the cod can no longer repopulate as they had for the previous 500 years. It is useful to see how this happened.

Fishery scientists set limits on the number of fish that can be caught, but do not have direct access to the number and weight of fish in the sea. Instead, they must rely upon indirect measurements (i.e., effects of the size of the fish population). One indicator comes from scientific surveys in which trawlers fish with standardized equipment to produce a controlled sample of the cod population. The carefully controlled nature of this data collection process makes it fairly robust: the causal relationship between the actual fish population and the data collected is relatively stable.[7] Such surveys are expensive to conduct, however, so they are used only sparingly.

Instead, fishery scientists largely use reports from fishermen of their commercial *catch per unit effort* (CPUE). CPUE is used to estimate the amount of fishable biomass by means of a model. Intuitively, it makes sense that if there are more fish, it should be easier to catch them; fewer fish, and it should be more difficult to catch them. This idea is captured in the simple model used by fishery scientists—$CPUE = q * fish$—that supports estimates of the fishable biomass given CPUE: $fish = \frac{CPUE}{q}$. The parameter q, commonly called the "catchability coefficient," was assumed to be fixed for any particular fish population.[8] That is, the scientists assumed that context-driven model change could not occur: the *CPUE—fish relationship*, whatever it might be, was assumed to be stable from year to year (though the values of *CPUE* and $fish$ obviously fluctuated).

Despite the model's simplicity and strong assumptions, it was quite accurate for many years: from 1962 to 1974, there seems to have been a stable target model (see Fig. 2)[9] that yielded accurate estimates of the actual fish population. After 1974, however, the relationship between *CPUE* and $fish$ changed as (i) new technologies were introduced; (ii) fishermen's knowledge of the area increased; and (iii) the decreasing cod population became sparsely distributed in dense pockets. As a result, *CPUE* was no longer a linear function of $fish$; that is, there had been context-driven model change, though the fishery scientists were unaware of this fact.[10] Again, the key issue here is not that the value of $fish$ changed over time; rather, the problem was that the very function connecting $fish$ and *CPUE* had changed in a fundamental way. The

[7] Of course, the fish population varies from year to year, but that is *not* model change as we understand it. The important point here is that the relevant causal/sampling *relationship* does not change.

[8] Some researchers have argued that, under certain circumstances, reducing the amount of fish may actually make it easier to catch them, and so q could potentially be negative. In such contexts, the $CPUE - fish$ relationship would be piecewise linear, rather than a single linear function.

[9] In reading Fig. 2, one must remember that $fish$ is an unobserved variable. The $fish$ numbers were retroactively recalculated *after* fishery scientists realized that there had been context-driven model change. They were not available to the scientists at the time.

[10] There actually were two scientific surveys during this time period, and they unsurprisingly contradicted the significantly higher estimated fish population. Scientists were unsure of how exactly to reconcile the conflicting measurements, so they decided simply to average the CPUE estimates with the survey data, which still grossly overestimated the population.

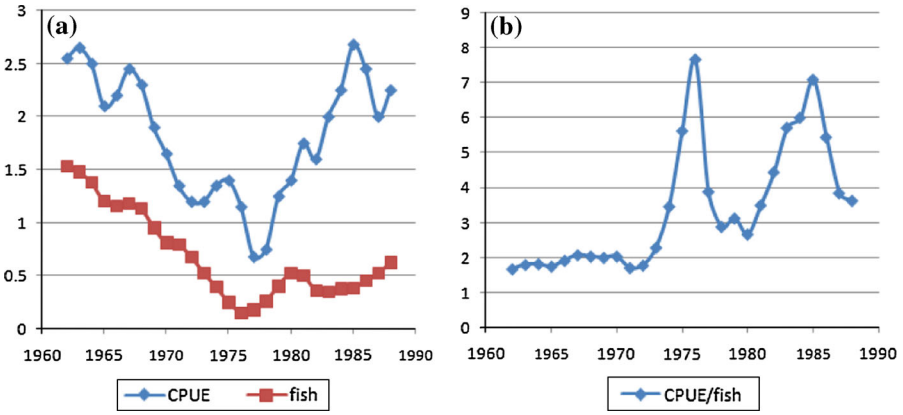

Fig. 2 **a** Fishable biomass compared to total CPUE from 1962 until 1988. (Adapted from McGuire 1997). **b** The ratio of CPUE to fishable biomass (the correct value for q)

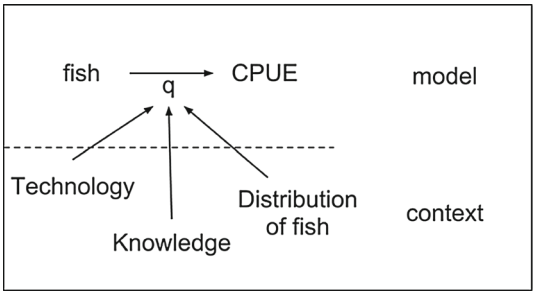

Fig. 3 Model used by fishery scientists and some of its context

1980's Total Allowable Catch was set using estimates derived from the old model, and the fishery collapsed shortly after.

In the language of Fig. 1, q $(= \Gamma)$ is a property of the relationship between $fish$ $(= A)$ and $CPUE$ $(= B)$. However, changes in elements of the context $(= C)$ such as technology, fishermen's knowledge, the distribution of fish, led to a change in q, and in fact to the very functional form between $fish$ and $CPUE$ (Fig. 2b). That is, there was context-driven model change (Fig. 3).

3.2 Context-driven model change in economics

In economics, context-driven model change has been recognized as a general class of phenomena that have been named *structural breaks*. Structural breaks are understood to be cases in which there is a shift in the (quasi-)objective *structure* of the world. More precisely, econometricians define a structural break to be a significant change in the underlying data generating process (Clements and Hendry 1999). Structural breaks thus differ from "mere" shocks in which values of variables are influenced or changed by factors outside of the system, but without substantively changing the relationships between the variables in the model. Many authors have argued that structural breaks

are pervasive and significant in macroeconomics (Clements and Hendry 1998, 1999; Hansen 2001).[11] A number of tools have been developed to combat this problem (see Perron 2006 for an overview) principally through model generalization: context-driven model change prompts the development of a more general model in which (causal) parameters can vary with time (e.g., locally stationary models, as in Dahlhaus, 1997; Dahlhaus and Polonik, 2009). This response can look like a model expansion, but the generalized models do not actually include the exogenous factors responsible for the context-driven model change; rather, they only model the resulting changes in the distributions for the original set of variables. Similarly, structural break detection and estimation (Davis et al. 2006; Davis and Rodriguez-Yam 2008) aims to model the shifts over time without necessarily explaining *why* the particular structural breaks occurred.

The "Great Moderation" is a well-known macroeconomic example of context-driven model change (Blanchard and Simon 2001). Many United States macroeconomic variables experienced a significant decrease in volatility from the mid 1980's until the 2007 financial crisis; this is referred to as the Great Moderation because it was a period of significant stability in economic phenomena. This was not simply a change in the values of the economic variables, but a fundamental change in the underlying relationships, both at a particular moment and over time. However, despite the size of the shift (it was the *Great* Moderation!), it went undetected for over a decade (Blanchard and Simon 2001). Moreover, this shift potentially had significant practical and societal implications, as some economists believed that it would provide the key to preventing economic depressions (Lucas et al. 2003).

4 Normative accounts of scientific inquiry

Context-driven model change poses a serious, but not insurmountable, problem for science. Scientists do, in practice, find ways to (partially) solve this problem, and so we should expect (or desire) that normative accounts of scientific inquiry provide guidance on reliably handling context-driven model change. One might hope that sophisticated inference and reasoning methods (e.g., various types of non-monotonic reasoning) would be suitable, but these methods all operate only on a given model. That model can represent very complicated between-variable relationships, but they nonetheless remain fixed for the duration of inference. Consider a simple case of non-monotonic reasoning in the $Switch$, $Light$, $Power$ case. Given $Switch = on$, one infers that $Light = on$, but that reasoning is defeasible: if one subsequently learns $Power = off$, then the $Light = on$ conclusion is retracted. This might appear to be a case in which the $Switch \rightarrow Light$ relationship changes, but *from the perspective of the reasoning method/model*, the relationship is constant-but-complicated. That is, the relationship appears to change only because we initially ignored the value of $Power$, but the method has no such luxury. In order to use $Power$ in its reasoning, it must have

[11] Hansen (2001) sums up the problem nicely: "Structural change is pervasive in economic time series relationships, and it can be quite perilous to ignore. Inferences about economic relationships can go astray, forecasts can be inaccurate, and policy recommendations can be misleading or worse." (p. 127)

included *Power* from the beginning (at least, implicitly), and so there is not actually any context-driven model change.

One might instead look to normative accounts of when scientists ought to choose one model rather than another. Unfortunately, all existing accounts essentially assume away context-driven model change.[12] The core problem is that all of these accounts assume that scientific inquiry is a process that aims to discover theories that are "fixed-for-all-time": for all of the accounts, the true (or accurate, or fruitful, or …) scientific theories represent a fixed target towards which scientists are or should be aiming. This assumption that the correct theories do not change over time has the impact of simply assuming away model change.[13] Space constraints preclude us from surveying all existing normative accounts of scientific inquiry, but we demonstrate this implicit assumption by considering two prominent, but quite different, traditions.

Karl Popper's falsifiability criterion (Popper 1963) treats the identification of false scientific theories as the key to normative inquiry: one should rigorously test the falsifiable (in principle) scientific theories, and when a theory fails some test, then it should be removed from the pool of potentially true theories. Over time, this process should converge towards a set containing only true theories, as the false ones will eventually be falsified and eliminated. Model change poses an obvious problem for this approach, as falsified theories are never reintroduced into the pool of potentially true theories. If a theory is falsified at time t, then it is no longer available at time $t + \Delta$, even though it could be correct at that later time.

In Deborah Mayo's related account of scientific inquiry based on severe testing (Mayo 1996, 1991, 1997; Mayo and Spanos 2006), scientists should subject their theory to a series of severe tests—tests that can not only falsify the theory, but also support positive belief in the theory if it passes the test.[14] The scientist's warrant in believing her theory depends on "the degree of severity with which a hypothesis H has passed a test" (Mayo and Spanos 2006, p. 328). Mayo's account is thus a strengthening of Popper's, as it shows how scientific inquiry can be a reliable process, in contrast with falsification.[15] At the same time, it shares with Popper the assumption that falsified theories—alternately, theories that have failed severe tests—are excluded from the subsequent possibility space. Both accounts could try to accommodate model change by allowing previously rejected theories back into the pool of possibilities, but this change threatens the long-run convergence guarantees that provide an important part of the accounts' normative justification.

A second example is the cluster of normative Bayesian accounts of scientific inquiry (Earman 1992; Howson and Urbach 1993), all of which are based on updat-

[12] Of course, it is certainly possible that the accounts discussed in this section could be adjusted to handle model change. We simply aim to show that existing ones do not currently accommodate it.

[13] One might object that accounts of scientific inquiry are about *theories*, not models (i.e., our focus), and that this makes a meaningful difference. However, it is not clear what the relevant difference would be. Moreover, the two accounts we consider below are both supposed to apply to more focused models, as well as broader theories.

[14] In statistical language, severe tests should have both a low false positive and a low false negative rate. Falsifying tests need only have a low false negative rate (where 'negative' means "theory is found false").

[15] Additionally, her account is relatively localist, as it advocates testing "small" claims (e.g., models) rather than entire scientific theories.

ing $P(H_i|E)$—the probability of each hypothesis given the evidence—using Bayes' rule whenever new evidence is observed. The standard (and natural) statement of an hypothesis H_i is in the form "Theory T_i is true for all times, space, contexts, etc." This yields as many hypotheses as scientific theories, and most Bayesian accounts use this sort of hypothesis space. In general, the hypothesis space (for a Bayesian account) cannot change during inquiry, and so the use of such universal hypotheses amounts to the implicit assumption that model change is impossible. In principle, Bayesian accounts could allow for model change by expanding the size of the hypothesis space to include ones that assert different theories at different times or spaces (e.g. "T_i at time t_1 & T_j at t_2 & …"). This leads to a combinatorial explosion of the hypothesis space, however, and so makes calculation of $P(H_i|E)$ computationally intractable. In addition, the normative grounding of Bayesian accounts becomes shaky when multiple hypotheses have identical likelihood functions (i.e., when H_i and H_j imply the same probabilities for possible data sequences), and that is significantly more likely with these more complex hypothesis spaces.

One might respond that these accounts are focused on "ideal" inquiry, and so on maximally general models that include all possibly relevant contextual factors (e.g., $Power$ or the various influences on the catchability coefficient q). The problem of context-driven model change simply does not arise if models explicitly incorporate every relevant contextual factor. The problem is, as we argued earlier, that maximal models with empty contexts would have to include everything in the past light-cone. Such models are certainly impossible for actual human scientists, and arguably impossible even in theory. In general, such normative accounts neither represent nor solve the problem of context-driven model change, even though actual scientists both recognize and (partially) solve it. We now show that scientists can (normatively) recognize and respond to context-driven model change, and even protect themselves against such model change. Unfortunately, this protection requires giving up on a standard normative desideratum of statistical methods for scientific inquiry.

5 A domain-general response

We desire normatively justifiable methods that can learn the appropriate target model, even when there is a possibility of context-driven model change. We also seek methods that can work in all sciences, so they should be domain-general. It is important to recognize that *any* such response will need to assume that we do not have large, frequent changes.[16] If the target model changes significantly at every moment in time, then there is no stable information for the learning method to use. We thus assume (as, we contend, all scientists do) that the real world phenomena that we want to model are sufficiently orderly and parsimonious (relative to the variables in our models) that they could potentially be discovered. This assumption is essentially a response to Hume's problem of induction: if the future can be arbitrarily different from the past, then no

[16] What counts as 'large' or 'frequent' will depend partly on our inference or learning method. Informally, we need to assume that the target model does not continue to change before we can identify it with the particular inference method in use.

reliable inference is possible, and so we must assume that there is *some* regularity. Importantly, however, we are making a significantly weaker assumption than one finds in standard normative accounts, as we do not assume that the target model is the same for all time.

Given this assumption of locally stable contexts, any domain-general methods that detect and respond to context-driven model change must instantiate this general schema:

1. *Recognize* context-driven model change by detecting a collection of anomalies
2. *Decide* whether to find the new target model by:
   (a) changing model components (e.g., parameters); or
   (b) changing the framework to include models with additional contextual elements
3. *Learn* the (appropriate) new target model

First, the method must recognize that model change has potentially occurred, either at that moment or soon after. If no model change is detected, then there is no reason to do anything different. Since we are focusing on domain-general methods, however, the only (somewhat) reliable signal of context-driven model change is a group of observations or phenomena that cannot be predicted or explained by a model M that had previously done a good job predicting and explaining the data (i.e., was plausibly the target model). Such observations are typically characterized by the scientific community as anomalies, precisely because they cannot be explained by the model that had previously been successful.[17]

Anomaly detection is problematic precisely because one should expect low-probability events to occur *even if* the underlying distribution is stable; a 1 in a million (or worse) chance will occur in an infinite stream of data with probability 1. Thus, one unusual datapoint (i.e., a single outlier) should not necessarily prompt suspicion of context change. An *anomaly collection* is a subsequence of data that deviates so strongly from the expected data that we can be sufficiently confident that it does not come from the same distribution as the previous data. Since models (in stable background conditions) imply stable observational data distributions, a change in the distribution implies a change in the target model.[18] Anomaly collections provide a strong signal that there is a new target model, but for most models, no finite amount of data will *guarantee* that the target model has shifted, though it might be exceptionally probable that a shift occurred.

Second, one must decide how to respond to the suspected context-driven model change. One approach is to choose a new model (hopefully, the actual new target model) from the same framework (i.e., approach 2(a)). For example, one might simply relearn certain parameters of the between-element relationships (e.g., linear coefficients). A different approach is to change the framework to include models that explicitly incorporate the contextual factors that drove the model change (i.e., approach 2(b)).

[17] Such anomalies are similar in spirit to Kuhnian anomalies, but are of course on a much smaller scale.

[18] The converse obviously does not hold: different models can produce the same data distributions, so context-driven model change does not necessarily lead to anomalous data. This is an instance of the general problem of underdetermination of models by observed data that affects *all* model inference methods, including ones that do not accommodate context-driven model changes.

In many cases, the old target model will then be a sub-model of the various possible models in the new framework.

These two approaches have different advantages and disadvantages. In practice, 2(b) is typically more difficult to implement than 2(a), since we have to determine the appropriate factors, incorporate them into the possible models, and determine all of the relationships they bear to factors that were in the old target model. There is arguably no domain-general way to implement 2(b), as the decision about which factors might have mattered will invariably depend on domain-specific information. At the same time, 2(b) has the significant virtue that the resulting target model is stable both pre- and post-context change; more precisely, the "context change" has been reinterpreted as simply a change in the values of modeled variables within a stable model. The decision about which approach to pursue will depend on weighting the practical advantages of 2(a) against the theoretical virtues of 2(b).

Approach 2(b) also has the practical drawback that learning a new target model (step 3, discussed below) will require data about the factors that were formerly in the context, but are now in the models. Because they were previously in the context, we are unlikely to have all of the data that we need to learn the new target model, so learning will be delayed while data collection occurs. In contrast, approach 2(a) uses the same variables that we already have, but recognizes that some parameters may have changed. We thus need to reduce the relative influence of data generated in contexts that are dissimilar from the current one, though we will frequently be able to reuse at least some of the previously collected data.

One standard way to reuse the data intelligently is through *downweighting*: treat the downweighted points as fractions of points (or ignore them entirely), as in a weighted average. Different datapoints can be independently downweighted, so the space of possible downweighting schemes is large. If one knows only that context change has occurred but nothing about the similarity between previous target models and the current one, then one could simply give all previous datapoints zero weight. If the current target model is close to previous ones, then previous datapoints should perhaps be downweighted only partially.[19] As a more interesting example, suppose one believes that there are alternating target models: C_1 holds for times $[0, t_1), [t_2, t_3), \ldots, [t_{2n}, t_{2n+1}), \ldots$ and C_2 holds for $[t_1, t_2), [t_3, t_4), \ldots, [t_{2n-1}, t_{2n}), \ldots$. In this case, one should dynamically adjust the downweighting over time—sometimes downweighting a datapoint and sometimes not—so that one is always learning with data from the current target model.

Third, the new target model must be learned from either the original framework or a new framework of expanded models, and using either downweighted data or data from an expanded set of variables. The details of this step will be determined by the particular learning method that one uses, typically the same one that was previously used to learn target models, perhaps with some small adjustments. For example, statistical methods—those that use statistical estimation to infer (frequency) distributional infor-

[19] It is tempting to think that one should always throw away all previous data whenever context-driven model change occurs. However, this strategy potentially throws away useful data that could lead one to the target model more rapidly, and also does not allow for the possibility of retracting an anomaly detection judgment.

mation about a set of defined variables—are widespread in the sciences (e.g., in a range of standard causal inference algorithms as in Spirtes et al., 2000; Pearl, 2000, or various sophisticated model selection/averaging methods), and can be adjusted to allow for the possibility of context-driven model change in relatively straightforward, and enlightening, ways. For example, we provide (in Anonymous, 2013) a causal learning algorithm that can respond in real-time to changes in the underlying causal structure. In particular, that method weakly dominates existing causal structure learning methods: it performs equally when the underlying causal structure does not change, and significantly outperforms previous methods when changes do occur, precisely because it can learn the new causal structure.

5.1 Diligence versus consistency

This protection against context-driven model change comes at a cost: responsiveness using approach 2(a) is incompatible with a standard methodological virtue, at least for a wide class of methods based on statistical estimation and distributions of observed data.[20] Model expansion (approach 2(b)) does avoid the following issues as it involves changing estimators "mid-stream," but it involves significant practical challenges and is not a domain-general response.

We first must define two desiderata for statistical estimators: consistency and diligence. Target models in frameworks of statistical models are identified by statistical estimation from observed data. One of the weakest (and so widely-assumed) virtues of a statistical estimator is *consistency*, also known as convergence in probability. Informally, an estimator is consistent just when, with probability 1, it outputs the target model given an infinite stream of data (under certain assumptions). That is, if there is a stable target model M for all time, then as the estimator is provided with more data from M, the probability that the method's "answer" is arbitrarily close to M approaches 1. It should be clear why this is a desirable property: estimators that cannot even weakly guarantee (in a probabilistic sense) that they reach the target model in the infinite limit arguably cannot be trusted on short-run, real-world data. Moreover, many other plausible virtues of estimators—e.g., almost sure convergence, convergence in quadratic mean, sure convergence—strictly imply consistency. Essentially any statistical estimator with a claim to being reliable in the long run is consistent.

The possibility of context-driven model change suggests a different methodological virtue, *diligence*. Suppose we have a model change event at time t, after which the target model is again stable. Informally, an estimator is diligent if there is some finite amount of data Δ (which can depend on the change but does *not* depend on t) such that the estimator will have a strictly positive probability of outputting the new target model after $t + \Delta$. In other words, no matter how much data we have previously seen,

[20] We conjecture that the general "consistency vs. diligence" tension that we discuss below arises for essentially all methods of scientific inference, but only have a formal proof for statistical estimators. It is an open research question whether this incompatibility can be proven for a broader class of methods, though we note that an enormous part of scientific inquiry consists in statistical estimation. In general, we suspect that part of the reason that normative accounts of science have ignored context-driven model change is precisely because they privilege consistency, and so cannot value diligence in the same way.

we have a chance of learning the new target model after a fixed (change-specific) amount of post-change data. Diligence implies that, regardless of the amount of prior data, we will not remain ignorant of a model change for arbitrarily long. Put more colloquially, the method diligently evaluates all of the data, no matter how much data it has previously seen. Given the practical and societal importance of context-driven model changes, it is a highly desirable property for our statistical estimators.

As diligence is a novel methodological virtue, it might be helpful to give a simple example. Suppose that we are measuring the value of a single, real-valued variable X, and our framework (i.e., set of models) consists of all possible Normal (i.e., Gaussian) distributions. For this framework, one inference task is to estimate the mean of X: $\mu(X)$. The simplest estimator of $\mu(X)$ is the average of all measurements, but this estimator is not diligent. If the mean changes from μ_1 to μ_2, then the length of time that this estimator will be "fooled" (i.e., be far from μ_2) depends partly on how much data it saw from the μ_1-distribution; if we have seen 1,000,000 datapoints from the μ_1 distribution, then the estimate will take longer to converge to μ_2 than if we had seen only 1,000 datapoints. There is no fixed length of time within which this estimator will respond to the change regardless of the amount of data seen previously.

In contrast, the estimator that returns the average of only the last 100 measurements *is* diligent, since it will respond to any change within 100 datapoints, regardless of how much data it saw before the change. If the mean changes from μ_1 to μ_2 at time t, then we are guaranteed to have a close estimate of μ_2 by $t + 100$, regardless of whether we saw 1,000 or 1,000,000 (or more) datapoints from the μ_1 distribution. If it is important to quickly detect changes in the distribution mean, then this type of protection could be quite valuable. Similar observations can be made about more complex statistical estimators.

Both consistency and diligence are desirable methodological virtues. Unfortunately, they are *incompatible* virtues: no statistical estimator can satisfy both.[21] To see why, we need one additional notion. Let an ϵ, δ-*error* occur whenever the probability that the output of method M is within δ of the target model is less than ϵ. That is, it is unlikely (occurs with probability less than ϵ) that the method is close (produces an estimate no further away than δ) to the target. A method M is *subject to arbitrary errors* if, for every model change, ϵ, δ, and n, there is some length r of initial data that leads M to make n many ϵ, δ-errors after that model change. In other words, however we want to characterize errors, M is subject to arbitrary errors when we can force it to make n many errors in a row (for any n) by presenting enough initial data. More colloquially, we can always find a way to force M to be fooled for arbitrarily long after the model change. Obviously, any method that is subject to arbitrary errors leads one to be quite vulnerable to the effects of future context-driven model changes.

[21] This statement is not quite right: it is possible for consistent estimators to be diligent, but only under very special conditions given in "Construction: diligence $\Rightarrow \neg$ arbitrary errors" in Appendix section. Roughly, an estimator can be both consistent and diligent only when every model in the framework is sufficiently far away from every other model, so that certain data *guarantee* that model change has occurred. Few realistic frameworks satisfy this condition, though many toy ones do. For example, a framework for deterministic data generating processes with no measurement error meets this condition, as can (sometimes) a framework for deterministic data generating processes with bounded measurement errors.

Unsurprisingly, diligent estimators are not subject to arbitrary errors; the whole point of that virtue is that the estimator adjusts within a known length of time. Perhaps surprisingly, though, *all* consistent estimators are subject to arbitrary errors (given a nontrivial framework, where 'nontrivial' is defined precisely in the Appendix). We can thus immediately prove:

Theorem *No statistical estimator for a (nontrivial) framework is both consistent and diligent.*[22]

We thus have a real, but insurmountable, problem for scientific methodology. Basic intuitions about reliability imply that we should use consistent estimators. The existence of context-driven model change, and the very real practical and societal impacts of it, imply that we should use diligent estimators. But for most, if not all, of the modeling problems faced by modern scientists, no estimator can satisfy both of these desires: every estimator must choose whether to converge to the stable target model (when it exists), or to diligently watch for potential model change.[23] Consistent estimators are stable-but-conservative: they find the right answer (when it exists) precisely by ignoring unusual events. Diligent estimators are responsive-but-volatile: they can rapidly adjust to a changing world, but only by sometimes changing unnecessarily. Of course, what we want are estimators that are stable-and-responsive, but the above theorem tells us that we must choose between them.

The choice about whether to use a consistent or diligent estimator in any particular context is a complex, situation-specific one. The two key risks in this type of scientific inquiry are (i) missing an actual context-driven model change for some length of time; and (ii) thinking that context-driven model change occurred when it actually did not. Consistent estimators tend to minimize mistakes of type (ii); diligent estimators tend to minimize mistakes of type (i). Thus, if the costs of one of these errors outweighs the other, then we can intelligently choose an estimator to minimize the more costly mistake. In fact, for particular estimators, one can sometimes derive the probabilities of each type of error as a function of the sample size, current target model, and size of change that is "meaningful." If we have such probabilities for multiple estimators, as well as quantitative estimates of the costs of each type of error, then the decision about which estimator to choose becomes a simple exercise in minimizing some standard cost-benefit function (e.g., minimizing expected total costs, perhaps after transforming the risk probabilities). It is thus sometimes possible to make a principled decision about whether to use a consistent or diligent estimator, but much depends on situation-specific details.

6 Conclusion

We have argued that the possibility of context-driven model change—changes of context that result in a change of target model—arises naturally from our inability to

[22] The proof and precise statements of the above notions are provided in the Appendix.

[23] Of course, there are estimators that are neither consistent nor diligent, but we ignore those here.

include absolutely every possibly relevant factor in our models, but presents a significant challenge for scientific practice. Such model change is not rare, not insignificant, and not obvious, but has been largely ignored in normative accounts of science. There are natural ways to adjust our scientific inference methods so that they are robust against the possibility of context-driven model change, but these adjustments come at a cost: learning methods that are diligent (and so suitably "protected" against model change) provably cannot be consistent, and so fail to have a key property of reliable inquiry. We cannot find the right answer both (a) quickly when the world changes; and (b) reliably when the world is stable. Instead, we must trade-off these desiderata based on a complex set of considerations.

Appendix 1: Notation

Let X represent a random sequence of data. Let X^t_B represent a random subsequence of length t of data generated from distribution B. Let $\mathbf{F}$ be a framework (in this case, a set of distributions). Let $M_{\mathbf{F}}$ be a method that takes a data sequence X as input and outputs a distribution $B \in \mathbf{F}$; we will typically drop the subscript $\mathbf{F}$ from M as we will be dealing with a single framework at a time. Concretely, $M[X^t_B] = O$ means that M outputs O after observing the sequence X^t_B. Let D be a distance metric over distributions (e.g., the Anderson-Darling test). Let $D_\delta(A, B)$ be shorthand for the following inequality: $D(A, B) < \delta$. Finally, let $[X, Y]$ denote the concatenation of sequence X with sequence Y.

Definition A distribution A is *absolutely continuous* with respect to another distribution B iff $\forall x\ P_B(x) = 0 \Rightarrow P_A(x) = 0$. That is, if B gives probability 0 to some event x, then A also gives probability 0 to that same event. Let $AC(A)$ be the set of distributions which are absolutely continuous with respect to A except for A itself.

Definition An estimator M is *consistent* if $\forall B \in \mathbf{F}\ \forall \delta > 0 \lim_{n\to\infty} P(D_\delta(M[X^t_B], B)) \to 1$. That is, for all distributions in the framework, the probability that M's output is arbitrarily close to the target distribution approaches 1 as the amount of data increases to infinity.

Definition An estimator M can be forced to make *arbitrary errors* if $\forall B_1 \in \mathbf{F}\ \forall B_2 \in AC(B_1) \cap \mathbf{F}\ \forall \delta, \epsilon > 0\ \forall n_2 \exists n_1 P(D_\delta(M[X^{n_1}_{B_1}, X^{n_2}_{B_2}], B_2)) \leq \epsilon$. That is, consider any distribution B_2 which is in the framework and is absolutely continuous with respect to B_1. Then for any amount of data n_2 from B_2, there is an amount of data n_1 from B_1 such that M's output will still be arbitrarily unlikely to be arbitrarily close to B_2 after seeing the $n_1 + n_2$ data.

Appendix 2: Lemma: consistency ⇒ arbitrary errors (within AC)

Proof Assume M is consistent. It suffices to show that:

$$\forall B_1 \in \mathbf{F}\ \forall n_2 > 0\ \forall B_2 \in AC(B_1) \cap \mathbf{F}\ \forall \delta > 0\ \forall \epsilon < 1\ \exists n_1 P(D_\delta(M[X^{n_1}_{B_1}, X^{n_2}_{B_2}], B_1)) > \epsilon$$

That is, even if we add a finite sequence of data drawn from B_2 to the end of any $X_{B_1}^{n_1}$ sequence, then there is some amount of B_1 data so that the estimator M still converges to B_1.

Choose arbitrary B_1, B_2 and n_2. Let S be the set of all events in the metric space that, if satisfied by $X_{B_2}^{n_2}$, would stop M from converging to B_1. That is, let S be the set of all events that, if satisfied by $X_{B_2}^{n_2}$, would entail the negation of:

$$\forall\delta > 0\, \forall\epsilon < 1\, \exists n_1 P(D_\delta(M[X_{B_1}^{n_1}, X_{B_2}^{n_2}], B_1)) > \epsilon$$

Since M is consistent for B_1, then $P(X_{B_1}^{n_1} \in S) = 0$. Since B_2 is absolutely continuous with respect to B_1, $P(X_{B_2}^{n_2} \in S) = 0$. As such, it is at most a probability 0 event that $X_{B_2}^{n_2}$ can take a value that prevents M from converging to B_1, so M will still converge in probability to B_1 over sequences of the form $[X_{B_1}^{n_1}, X_{B_2}^{n_2}]$. □

Appendix 3: Construction: diligence ⇒ ¬ arbitrary errors

We construct the formal definition of diligence from that of "arbitrary errors" (AE) in a way that makes it clear that diligent methods are not subject to arbitrary errors. The negation of AE is:

$$\exists B_1 \in \mathbf{F}\, \exists B_2 \in AC(B_1) \cap \mathbf{F}\, \exists\delta > 0\, \exists\epsilon < 1\, \exists n_2 \forall n_1 P(D_\delta(M[X_{B_1}^{n_1}, X_{B_2}^{n_2}], B_2)) > \epsilon$$

This condition is, however, insufficiently weak to capture diligence, as we want to avoid such errors for *all* pairs of distributions in the framework, not just for some absolutely continuous pair. We thus strengthen the negation of AE by turning the two leading existential quantifiers into universal quantifiers and extending the domain of the universal quantifier over B_2 to include those distributions which are not absolutely continuous with respect to B_1:

Definition An estimator M is *diligent* if

$$\forall B_1 \in \mathbf{F}\, \forall B_2 \in \mathbf{F}\backslash B_1\, \forall\delta > 0\, \exists\epsilon > 0\, \exists n_2 \forall n_1 P(D_\delta(M[X_{B_1}^{n_1}, X_{B_2}^{n_2}], B_2)) > \epsilon.$$

That is, for any pair of distributions in the framework, there is an amount of data n_2 from B_2 such that M's output will be arbitrarily close to B_2 with positive probability after seeing $n_1 + n_2$ data, for *any* amount of data n_1 from B_1.

Definition A framework $\mathbf{F}$ is *nontrivial* iff there exists some $B \in \mathbf{F}$ such that $AC(B) \cap \mathbf{F} \neq \emptyset$.

Clearly, diligence implies the negation of AE for all nontrivial frameworks. We thus have the key theorem for this paper:

Theorem *No statistical estimator for a (nontrivial) framework is both consistent and diligent.*

Proof Assume M is both consistent and diligent. Its consistency implies that AE holds for it. Its diligence, along with the nontriviality of the framework, implies that ¬AE holds for it. Contradiction, and so no M can be both consistent and diligent for a nontrivial framework. □

Appendix 4: Generalizing diligence

A natural generalization of diligence yields a novel methodological virtue: Uniform Diligence. Uniform diligence is a strengthening of regular (pointwise) diligence in the same way that uniform consistency is a strengthening of pointwise consistency. Instead of requiring only that, for each B_1, B_2 and δ, there be some n_2, Uniform Diligence requires that there be some n_2 which works for *all* such combinations.

Definition An estimator M is *uniformly diligent* if

$$\exists n_2 \forall B_1 \in \mathbf{F}\, \forall B_2 \in \mathbf{F} \backslash B_1\, \forall \delta > 0\, \exists \epsilon > 0\, \forall n_1 P(D_\delta(M[X^{n_1}_{B_1}, X^{n_2}_{B_2}], B_2)) > \epsilon.$$

Obviously, consistency and uniform diligence are also incompatible, as the latter is a strengthening of diligence. The following chart shows three different ways of ordering the quantifiers in the definition of Diligence, producing methodological virtues of varying strength. The weakest, Responsiveness, is not incompatible with consistency. For space and clarity, **B** is used in place of $\forall B_1 \in \mathbf{F}\, \forall B_2 \in \mathbf{F} \backslash B_1\, \forall \delta > 0\, \exists \epsilon > 0$.

Responsiveness	Diligence	Uniform diligence
$\mathbf{B} \forall n_1 \exists n_2$	$\mathbf{B} \exists n_2 \forall n_1$	$\exists n_2 \mathbf{B} \forall n_1$

References

Blanchard, O. J., & Simon, J. A. (2001). The long and large decline in US output volatility. *Brookings Papers on Economic Activity*, *135–164*, 2001.

Carpenter, S. R., Cole, J. J., Pace, M. L., Batt, R., Brock, W. A., Cline, T., et al. (2011). Early warnings of regime shofts: A whole-ecosystem experiment. *Science*, *332*, 1079–1082.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford: Oxford University Press.

Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge: Cambridge University Press.

Clements, M. P., & Hendry, D. F. (1998). *Forecasting economic time series*. Cambridge: Cambridge University Press.

Clements, M. P., & Hendry, D. F. (1999). *Forecasting non-stationary economic time series*. Cambridge: MIT Press.

Cooper, R. M., & Zubek, J. P. (1958). Effects of enriched and restricted early environments on the learning ability of bright and dull rats. *Canadian Journal of Psychology*, *12*, 159–164.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics*, *25*(1), 1–37.

Dahlhaus, R., & Polonik, W. (2009). Empirical spectral processes for locally stationary time series. *Bernoulli*, *15*(1), 1–39.

Davies, J., & Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews*, *74*(3), 417–433.

Davis, R. A., Lee, T., & Rodriguez-Yam, G. (2006). Structural break estimation for nonstationary time series models. *Journal of American Statistical Association*, *101*, 229–239.

Davis, R. A., & Rodriguez-Yam, G. (2008). Break detection for a class of nonlinear time series models. *Journal of Time Series Analysis*, *29*, 834–867.

Duffy, M. A., & Sivars-Becker, L. (2007). Rapid evolution and ecological hostparasite dynamics. *Ecology Letters*, *10*, 44–53.

Earman, J. (1992). *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge: The MIT Press.

Edmonds, B. (2007). The practical modelling of context-dependent causal processes. *Chemistry and Biodiversity*, *4*, 2386–2395.

Eells, E. (1991). *Probabilistic causality*. Cambridge: Cambridge University Press.

Estes, J. A. (2011). Trophic downgrading of planet earth. *Science*, *333*, 301–306.

Finlayson, A. C. (1994). *Fishing for truth: A sociological analysis of northern cod stock assessments from 1977–1990*. Institute of Social and Economic Research, Memorial University of Newfoundland.

Glymour, B. (2008). Stable models and causal explanation in evolutionary biology. *British Journal for the Philosophy of Science*, *59*, 835–855.

Glymour, B. (2011). Modeling environments: Interactive causation and adaptations to environmental conditions. *Philosophy of Science*, *78*(3), 448–471.

Hansen, B. E. (2001). The new econometrics of structural change: dating breaks in US labor productivity. *The Journal of Economic Perspectives*, *15*(4), 117–128.

Howson, C., Urbach, P. (1993). Scientific reasoning: The Bayesian approach. Open Court, second edition, 1993. Original work published 1989.

Kurlansky, M. (1997). *Cod: A biography of the fish that changed the world*. New York: Walker and Company.

Lucas, R. E., Mcgrattan, E., Phelan, C., Prescott, E., Rossi-hansberg, E., Sargent, T., et al. (2003). Macroeconomic priorities. *American Economic Review*, *93*, 1–14.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science*, *58*, 523–552.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, D. G. (1997). Severe tests, arguing from error, and methodological underdetermination. *Philosophical Studies*, *86*, 243–266.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a neymanpearson philosophy of induction. *British Journal for the Philosophy of Science*, *57*, 323–357.

McGuire, T. (1997). The last northern cod. *Journal of Political Ecology*, *4*, 41–54.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Perron, P. (2006). Dealing with structural breaks. In K. Patterson & T. C. Mills (Eds.), *Palgrave handbook of econometrics, Vol. 1* (pp. 278–352). Basingstoke: Palgrave Macmillan.

Popper, K. (1963). *Conjectures and refutations Vol.28*. London: Routledge.

Sarkar, S., & Fuller, T. (2003). Generalized norms of reaction for ecological developmental biology. *Evolution and Development*, *5*, 106–115.

Scheffer, M., & Carpenter, S. R. (2003). Catastrophic regime shifts in ecosystems: Linking theory to observation. *Trends in Ecology and Evolution*, *18*, 648–656.

Scheiner, S. M. (1993). Genetics and evolution of phenotypic plasticity. *Annual Review of Ecology and Systematics*, *24*, 35–68.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge: MIT Press.

Yoshida, T., Ellner, S. P., Jones, L. E., Bohannan, B. J. M., & Lenski, R. E. (2007). Cryptic population dynamics: Rapid evolution masks trophic interactions. *PLoS Biology*, *5*, 1868–1879.

Yoshida, T., Jones, L. E., Ellner, S. P., Fussmann, G. F., & Hairston, N. G, Jr. (2003). Rapid evolution drives ecological dynamics in a predator-prey system. *Nature*, *424*, 303–306.