How to do things with theory
The instrumental role of auxiliary hypotheses in testing

Corey Dethier

[Preprint; please see here for the final version.]

Abstract

Pierre Duhem's influential argument for holism relies on a view of the role that background theory plays in testing: according to this still common account of "auxiliary hypotheses," elements of background theory serve as truth-apt premises in arguments for or against a hypothesis. I argue that this view is mistaken. Rather than serving as truth-apt premises in arguments, auxiliary hypotheses are employed as (reliability-apt) "epistemic tools": instruments that perform specific tasks in connecting our theoretical questions with the world but that are not (or not usually) premises in arguments. On the resulting picture, the acceptability of an auxiliary hypothesis depends not on its truth but on contextual factors such as the task or purpose it is put to and the other tools employed alongside it.

> [I]n saying these words, we are *doing* something ... rather than *reporting* something.
>
> Austin (1962, 13)

0 Introduction

Pierre Duhem's incredibly influential *The Aim and Structure of Physical Theory* (1914/1951) is usually and rightly remembered for its holism. Central to the argument for holism—and perhaps more influential than the holist conclusion itself—is a view of the role that background theory plays in testing. For Duhem, theories or models other than the hypothesis that we aim to test serve primarily as premises in the derivation of consequences. These consequences are then compared to the world; since all the premises have the same logical relationship to the consequence, either all of the premises pass as a conjunction, or they fail as the same. Hence holism.

Call these background theories and models "auxiliary hypotheses." The central thesis of this paper is that the Duhemian view of the role that auxiliary hypotheses play in testing is mistaken. The theories, models, and analogies employed in testing a hypothesis rarely serve as premises. Instead, they serve as "epistemic tools": instruments that perform specific tasks in connecting our theoretical questions with the world but that are not (or not usually) premises in arguments. For the purposes of testing, therefore, it's largely irrelevant whether these theories or models represent truthfully. What we care about is whether they reliably perform their task.

The structure of the argument is as follows. I begin by outlining the Duhemian account of auxiliary hypotheses (§1), discuss two counterexamples to the view (§2), and argue that the failure of the account should be attributed to the fact that auxiliary hypotheses are not truth-apt representations but reliability-apt tools (§3). I end the paper with a more thorough sketch of this alternative, which I call the "epistemic tool view." On this alternative picture, the acceptability of an auxiliary hypothesis depends not on its truth but on contextual factors such as the task or purpose it is put to and the other tools employed alongside it (§4).

A final note: I see this paper as fitting within a broader tradition of arguments—some directed at versions of holism, some not—to the effect that logical relationships between sentences or propositions are not the sole exemplar or paradigm of reasoning in the sciences. As I see things, this tradition dates at least as far back as Hacking's discussion of microscopes (Hacking 1983)—though one could trace it even further back to Suppes (1962)—and includes a large number of the major works on general philosophy of science from the turn of the century.[1] The purpose of this essay is to argue that much of what goes into testing our best scientific theories is more like a microscope than it is like a premise, and to spell out precisely where and why someone convinced by this claim should depart from the older and still largely entrenched tradition.

1 Duhem's argument for holism

For the purposes of this paper, I take Duhem's argument for holism—or what's sometimes termed the "(non-)separability thesis" (Ariew 2018; Quinn 1974)—to be the one given in the first two sections of chapter 6 of *Aim and Structure*. I reconstruct it as follows:

(P1) A hypothesis is tested when (read: iff) a consequence derived from it is compared with the world (Duhem 1914/1951, 180).

(P2) All consequences derived from a theory are derived from the auxiliary hypotheses in the same sense as they are derived from the hypothesis we are interested in testing (Duhem 1914/1951, 182).

(P3) In actual testing scenarios, "science ... taken as a whole" is required as an auxiliary hypothesis in the derivation (Duhem 1914/1951, 187).

(C) In actual testing scenarios, "science ... taken as a whole" is tested.

The least interesting premise is (P2), which expresses a simple fact about arguments: every premise shares the same logical relationship with the conclusion. (P1) and (P3) are more contentious. The former expresses

[1] See, for example, Azzouni (2000), Longino (1990), Mayo (1996), Morgan & M. Morrison (1999), J. Norton (2003), Wilson (2006), and Woodward (2003).

a commitment to a form of hypothetico-deductivism: testing is the comparison of the logical consequences of a theory or hypothesis with the world (J. Morrison 2010, 344-46). I take it that the "is tested when" in (P1) should be interpreted as a claim about necessary and sufficient conditions. (P3), in turn, is something that looks like an empirical claim about the nature of science, namely that the theories that we test are highly interconnected (J. Morrison 2010, 339).

Together, (P1)-(P3) imply that "science ... taken as a whole" is what is tested. As such, in each case in which a hypothesis h is tested, the auxiliary hypotheses employed in deriving consequences from h are also tested. And thus, if the auxiliary hypotheses employed entail consequences that are false, h and the auxiliary hypotheses fail the test together.

Much of this argument is widely rejected today; both (P1) and (P3) face substantial criticism. Increases in sensitivity towards the methodology actually employed in science have seriously challenged Duhem's strong commitment to an H-D account (e.g. Longino 1990; Morgan & M. Morrison 1999), probalistic models of testing have presented alternatives to Duhem's strictly deductivist picture (e.g. Strevens 2001), and a number of philosophers have argued that any interesting version of (P3) is far too extreme to match actual practice (e.g. Azzouni 2000; Sober 1999). Nevertheless, some aspects of the account are still widely held. In particular, the view that the primary role of auxiliary hypotheses is as premises in arguments remains common—even if many philosophers have dropped Duhem's position that they serve as premises in specifically deductive arguments. For instance, the difference between Duhemain and Bayesian approaches is that the latter allow for (a) probabilistic relationships between evidence and the "theory as a whole" and (b) differences in the prior probabilities assigned to the elements of the theory as a whole (see Strevens 2001). But they still treat auxiliary hypotheses as propositions that are truth-apt and evaluable on that basis; we've merely switched from "is it true?" to "how likely is it to be true?"

Call the view that auxiliary hypotheses serve as truth-apt premises in testing the "Duhemian" account of auxiliary hypotheses.[2] The central implication of this account that I'll be concerned with in this paper is that the falsity of any given auxiliary hypothesis—the failure of one of the models employed in part of a test to have true assumptions, for example—renders the argument for a hypothesis unsound and thus undermines an entire successful test. This conclusion is clearly part of Duhem's view: if the theory as a whole is falsified, then the hypothesis cannot have passed the test. It's also an implication of the Bayesian view: if e only changes the the probability of h given a (or the conjunction of h and a), then the falsity of a renders e irrelevant to h. This aspect of the view also seems to be widely held within philosophy of science more broadly. Consider the literature on robustness. What Woodward (2006) calls "inferential robustness" occurs when members

[2] I'm not interested in diving into questions of what Duhem really meant, and will generally refer to the schematized "Duhemian" from now on; for a discussion of Duhem's intended argument, see Ariew (2018).

of a set of auxiliary hypotheses are interchangeable in the sense that a given piece of data will confirm or disconfirm a given hypothesis regardless of with auxiliary hypothesis is employed. Philosophers have largely been skeptical of the confirmational value of inferential robustness on the grounds that the interchangeable auxiliary hypotheses are usually mutually exclusive, meaning that at most one of them can be true and thus at most one of the arguments for the hypothesis has true premises (see Cartwright 1991; Houkes & Vaesen 2012; Odenbaugh & Alexandrova 2011; Orzack & Sober 1993; Woodward 2006). It's widely accepted in this literature that the degree of confirmation in a case of inferential robustness is proportional to the probability that one of the set of auxiliary hypotheses is true, meaning that—as noted by Odenbaugh & Alexandrova (2011)—if we know that none of them are true, inferential robustness provides no confirmation. So, much like Duhem, these philosophers are committed to the view that the falsity of an auxiliary hypothesis prevents the hypothesis under test from being confirmed.

Even if Duhemian varities of both holism and hypothetico-deductivism have largely been abandoned, the central Duhemian view of auxiliary hypotheses remains. For both Duhem and many contemporary authors, auxiliary hypotheses are truth-apt representations of the world. The primary use of such representations is as premises in arguments (whether deductive or not) that fail if one of the premises is false. My view is that this Duhemian account is mistaken; rather than truth-apt representations, auxiliary hypotheses are reliability-apt tools. In the next section, I discuss a pair of counterexamples to the account: in both of cases, (a) the hypothesis passes the test and (b) the auxiliary hypotheses are false.

2 Counterexamples to the Duhemian account

2.1 Pendula and testing universal gravity

Imagine a spinning spheroid made of liquid: as it spins faster, more and more of the liquid will tend towards the equator—the spheroid will squash in on itself—due to the fact that the gravitational force acting on a molecule at the pole will not be offset by centrifugal effects, while the gravitational force acting on a molecule on the equator will be. The degree of curvature of the spheroid will therefore depend on the ratio of these two effects such that the stronger the gravitational effect relative to the centrifugal one, the more perfectly spherical the spheroid. As a consequence, different theories of gravity will have different implications for the curvature of a spheroid in these circumstances, and thus different implications for the curvature of the earth. As Schliesser & Smith (forthcoming) have thoroughly documented, testing the curvature of the earth using pendula thus served as a major test of Newtonian gravity against its only real 17th and 18th century competitor—Huygens' theory of inverse-square celestial (i.e., non-universal) gravity.

Ultimately, the evidence that Huygens and Newton were after comes in the form of lengths of string. More specifically, they were looking for the differences in the length of a seconds-pendulum, or a pendulum whose period is two seconds—one second in each direction—at different locations on the globe. Why lengths of string? First, as just discussed, different theories of gravity have different implications for the curvature of the earth. The lower the ratio of gravitational to centrifugal effects, the greater the difference between the distance from the center of mass to a point on the equator and the distance from the center of mass to the poles. Second, these differences in distances have implications for the surface gravity at different points on the earth: the larger the difference in distance, the larger the difference in the strength of surface gravity. And differences in the strength of surface gravity, finally, have implications for the period of a pendulum, and thus for the length of string necessary to create a seconds-pendulum.

Historically speaking, the tests of the curvature of the earth don't quite qualify as an *experimentum crusis*: due to errors in testing, the data produced in early expeditions were too varied to simply settle the issue, and further theoretical developments showed that Huygens' account was unlikely to succeed before better empirical evidence was available (Schliesser & Smith forthcoming). Nevertheless, the test is well set up to be a crucial experiment in that the sole controversial premise under test is universal gravity. As Duhem emphasizes, however, it would be erroneous to see the hypothesis of universal gravity as the *only* theoretical element involved in the test. Auxiliary hypotheses are required in the derivation of the relevant predictions and it would always be possible for either Newton or Huygens to reject one of these assumptions rather than accept the failure of their account of gravitation. For our purposes, we can focus on just one of these assumptions, the model of the relationship between acceleration (a) and the observable features of a pendulum, namely length (l) and period (T), which is given by the following equation:

$$a = 4\pi^2 \frac{l}{T^2} \tag{1}$$

The use of the pendulum as an instrument to measure acceleration requires the use of (1): it's only with the latter that we can use the data delivered by the pendulum as a test of our competing gravity hypotheses (M. Morrison 2015, chapter 6).

For the philosopher committed to the Duhemian view of auxiliary hypotheses, therefore, our test confirms universal gravity if and only if (1) is true—that is, if and only if it can serve as a premise in an argument for universal gravity (whether an H-D argument or a more inductive one). Unfortunately, (1) is false; it's an idealization in that it fails to account for amplitude, which does have an effect on the relationship between acceleration and period. What is true is that this effect is negligible when the amplitude of the pendulum is very small. Practically speaking, the result is that the experimenter must take precautions to ensure that the amplitude of the pendulum actually used in an experiment is not too large (as Huygens in fact directed his experimenters to do; see Schliesser & Smith forthcoming). I take it, therefore, that the test provides good reason to accept universal gravity even though (1) is false. As such, this case provides a counterexample to the Duhemian account, at least at face value: the test can confirm universal gravity even though the auxiliary hypotheses employed in testing it are false.

2.2 Hardy-Weinberg models and heterozygote advantage

For alleles in which a heterozygote exhibits different features from a homozygote, it is not always evolutionary advantageous to be homozygous.[3] A situation in which the fittest individual is the heterozygote is known as a "heterozygote advantage"—the most famous case being sickle-cell anemia—and such an advantage can serve to maintain polymorphism at a particular locus even if one of the alleles acting at the locus dominates the other(s).

Testing for heterozygote advantage at a particular allele (like much else in population genetics) begins with a Hardy-Weinberg (HW) model.[4] Consider some population in which two alleles (A_1 and A_2) compete at a single locus such that p is the frequency of A_1 in the population for the current generation and q the frequency of A_2. Let $G_{A_1A_1}$ indicate the percentage of individuals in the current generation that are homozygous for A_1 (similarly for $G_{A_1A_2}$ and $G_{A_2A_2}$) and $G'_{A_1A_1}$ the percentage of individuals in the next generation homozygous for A_1 (as above). When applied to such a population, a HW model returns the expected values for the genotypes of individuals in subsequent generations: $G'_{A_1A_1} = p^2$, $G'_{A_2A_2} = q^2$, and $G'_{A_1A_2} = 2pq$.

The allele distributions returned by the HW model will only accurately reflect the real population under certain conditions. Most famously, the derivation only holds if the population in question is infinite, but a number of other conditions must also be met: there must be no mutation, no differential fitness, no organization of mating within the population (e.g., A_1A_1s can't prefer to mate with other A_1A_1s), and no genetic exchange with other populations or between generations (see Tempelton 2006, chapter 2). The result is that no actual population meets all these criteria; any HW model of a population is bound to misrepresent its target population in a myriad of ways. The best we can hope for is effective approximation: for the purposes of determining the distribution of alleles in the next generation, the difference between (say) the size of the population in an HW model and the actual population size does not make a difference.

When testing for heterozygote advantage, however, the situation is even more complicated. If there is a

[3] This may seem obvious, but the issue of just how heterozygous populations are was a fraught one for mid-20th century biology; see, e.g., Lewontin (1974).

[4] Since I began writing this essay, I've become aware that Elgin (2017) employs HW models to make a similar point to the one made here. For an influential philosophical analysis of HW models, see Sober (1984).

heterozygote advantage, after all, then by hypothesis the conditions under which the HW model effectively approximates the world don't obtain, as there's selection acting on the actual population. Nevertheless, any test for heterozygote advantage depends on the application of the HW model to the population of interest: what provides evidence for both the existence and strength of the heterozygote advantage is the deviation between the allele distribution predicted by the HW model and the actual allele distribution. Effectively, the HW model acts as a (theoretical) control population: it tells us what the allele distribution *would* be if there were no selection and thus no advantage, and it is the difference between the real population and the control that informs us that selection is in fact occurring.[5]

So: in testing for heterozygote advantage, we require the use of a model whose basic assumptions are guaranteed to be false in any real case. Without this model, we cannot derive predictions from either the heterozygote advantage hypothesis or its negation. In spite of the fact that it relies on false assumptions, the model allows us to make clear discriminations between the two hypotheses: if there's a heterozygote advantage, then $G'_{A_1A_2}$ will be substantially larger than predicted by the HW model, and will remain so in subsequent generations (i.e., for $G''_{A_1A_2}$); if there's no advantage, we shouldn't see this effect. As such, this case provides a second counter-example to the Duhemian view: we can test for heterozygote advantage even though the auxiliary hypotheses are false.

3 Diagnosing the problem with Duhem's argument

The cases just given provide counterexamples to the Duhemian account of auxiliary hypotheses: they're cases in which the auxiliary hypotheses are false but nevertheless there's a successful test of the hypothesis. The conclusion that we should draw is that the Duhemian is incorrect to see auxiliary hypotheses as truth-apt representations. In this section, I argue that while it's true that for the purposes of *derivation*, auxiliary hypotheses must be treated as truth-apt representations—logical entailments only hold between truth-bearers—*testing* a hypothesis is a different activity than deriving consequences from it (alternatively: "confirming" is not "predicting"; see J. Morrison 2010). For the purposes of testing, the role played by auxiliary hypotheses is more akin to that of an instrument, which is more or less reliable, rather than that of a premise, which is more or less true.

Consider a thermometer. What a thermometer does—perhaps better, what the scale on a thermometer does—is allow us to use the length of a column of fluid as a proxy for temperature. Thermometers can be better or worse at this job, in the sense that they can be more or less accurate and more or less precise,

[5] Of course, this description is idealized. The inclusion of other stochastic processes such as drift or complicating factors such as linkage disequilibrium mean that such inferences will necessarily be statistical. But the simple description suffices for our point.

but every decent thermometer will have some contexts in which it is relatively reliable and some contexts in which it isn't. If we get too hot, for example, the casing of any existing thermometer will simply melt, in which case it will no longer be able to perform its role. Most thermometers are much more limited. A thermometer designed to measure the temperature outside may only be scaled for a range of 200° F and for normal purposes only needs to be reliable within a more limited range than that—a range of 120° F could be expected to be sufficient every day in most years in the northern United States. Notice that the thermometer isn't reliable in a particular context *in spite* of its unreliability in another context. On the contrary, the failure of a thermometer to perform reliably on the surface of the sun has nothing whatsoever to do with its ability to reliably tell me whether it is below freezing outside. For the purposes of testing a hypothesis about whether it's freezing outside, then, there's no point in asking whether or not the thermometer would be reliable in radically different situations.

Importantly, (1) plays exactly the same role in testing hypotheses about the nature of gravity that the thermometer plays in testing hypotheses about whether it's freezing outside. Just as the thermometer allows us to use the length of a column of fluid as a proxy for temperature, (1) allows us to use the length of a pendulum string as a proxy for acceleration. And just as the thermometer can deliver inaccurate results in a distant context without impugning its reliability for everyday purposes, our pendulum equation can deliver inaccurate results in high-amplitude cases without impugning its reliability when the experiments are appropriately constrained so that the amplitude remains low. What matters in both cases is the *in-context reliability*—or what Parker (2009a) calls the adequacy-for-purpose—of the translation from the observed quantity to the quantity that it serves as a proxy for.

Similar comments apply in the case of the HW model. As with a thermometer, there's little sense in asking whether a control on an experiment is "true." Instead, what the control allows us to do is identify and isolate the phenomenon that we are interested in measuring: by comparing the recovery rate in the treated population and that in the control we can determine whether or not a drug is effective.[6] The recovery rate of the treated population alone is meaningless—and just like the thermometer can play the role of translating more or less reliably, the control may be more or less reliable as a tool for isolation. Notice that in order to perform this controlling role relative to a heterozygote advantage hypothesis, the Hardy-Weinberg model does not need to be interpreted as giving us the truth about what the allele distribution *would* be in the no-advantage case: all that's necessary is that the model needs to be a reliable guide to how likely the observed distribution is under such conditions.

The argument and cases just given serve to show that while the truth of auxiliary hypotheses may be sufficient for reliability, it isn't necessary: if it was, then neither the pendulum equation nor the Hardy-

[6] Bokulich (forthcoming) refers to a similar function as "subtracting"; see also S. D. Norton & Suppe (2001).

Weinberg model (nor, arguably, the thermometer, which isn't anything like a truth-bearer) could play the roles that they do. So what then is meant by reliability? Intuitively, the pendulum equation given above is reliable if and only if it *preserves information*: when we feed it good data on the length of a seconds-pendulum, it correctly determines the actual acceleration to the desired level of accuracy.[7] By contrast, the HW model is reliable if and only if it returns accurate values for the probability of the observed readings given a no-advantage hypothesis. The relevant notion of reliability, therefore, must be relativized to the task or question at hand—and in fact, the context in which that question is asked. In the context of the first example, the reliability of a model depends on the experimental setup (is the amplitude high or low?) and the level of accuracy needed to distinguish between the competing hypotheses. Similarly, a point-mass model of the solar system can be highly reliable with regards to the motions of the planets while being extremely unreliable with regards to their density. As such, the notion of reliability we want is something like: given a question, accurate data, and context, a model (or theory) is reliable iff the probability that it delivers the correct answer to the question is above a given threshold.

There's more developing of this positive picture to do, and I turn to that project in the next section. First, however, allow me to sum up the lessons of the counterexamples. In the context of testing, auxiliary hypotheses are often employed even though they're known to be false, strictly speaking. In these cases, it doesn't even seem accurate to say that scientists *treat* these models and theories as true. It's not the case, for example, that physicists pretend that (1) is true even though they know it is false. Instead, they treat the equation as reliable or "close enough" under certain conditions—with the implication (noted above) that the experimenter must ensure that these conditions of reliability hold. It's hard to make sense of this implication if we interpret them as pretending that these auxiliary hypotheses are true: if we're assuming that (1) is true, why insist on restricting the amplitudes of pendula? The positive view of testing that emerges is one in which auxiliary hypotheses have been transformed from premises into *epistemic tools* (or "inferential tools"; see Currie 2018, 263): their primary role in testing is not that of truth-apt premises in an argument or derivation but rather as reliability-apt instruments for for achieving specific tasks, such as an empirical phenomenon or translating observable quantities to unobservable ones. The problem with the Duhemian account of auxiliary hypotheses, therefore, is just that it conflates what's required for an auxiliary to serve as a premise with that's required for it to play one of these instrumental roles. While we might consider a premise a type of tool—a tool for carrying out derivations—it's clearly neither the only type of tool nor particularly representative of the category more generally.

[7] Compare Filion & Moir (2018, 739), who employ a related notion of reliability in a more technical context.

4 Auxiliary hypotheses as epistemic tools

I've argued that there are counterexamples to the Duhemian account of auxiliary hypotheses and that these counterexamples can be traced to a conflation of the role that auxiliary hypotheses play in testing with the role that they play in derivations. In this last section, I try to say something more concrete about the alternative: what does it mean to say that auxiliary hypotheses are more like tools than they are like premises? I identify three aspects of "epistemic tool" view that I'm defending. First, that the reliability (unlike the truth) of auxiliary hypotheses is relative to a context (§4.1); second, that this context sensitivity has importantly different holist implications than found in Duhem's (P3) (§4.2); third, that physical instruments and background theory play the same role in confirming tested result (§4.3).

4.1 The context-sensitivity of reliability

The first and most central aspect of the epistemic tool view is that the theories and models employed as auxiliary hypotheses are employed to accomplish distinct *tasks* or *functions*; as Parker (2009a) argues in the context of climate models, their reliability or adequacy is relative to a particular purpose. In the pendulum example of §2.1, for instance, (1) is employed to translate values for the length of string into values for acceleration; as discussed in §2.2, the HW model is used to isolate the phenomenon or as control on the experiment. Theories and models perform other tasks as well. In climate science—which provides some of our best examples of what Edwards (2010) calls "model-data symbiosis"—models are employed to interpolate (Edwards 2010, chapter 8), correct (Bokulich forthcoming; Lloyd 2012), re-analyze (Edwards 2010, chapter 12; Parker 2016), and even generate data (Winsberg 2018, chapter 4). There's no guarantee that a model that can reliably perform one of these tasks will even be equipped to perform another, let alone to do so reliably. Just as a thermometer simply doesn't provide us with the ability to measure the direction of the earth's magnetic field, a model that helps us identify unwanted causal influences (e.g., tree clearing that exposes a thermometer to more sun during the day) will not necessarily be able to provide us with estimates of the temperature in the middle of the Atlantic based on readings on the coasts.

In the context of testing, then, what we demand of a particular auxiliary hypothesis is that it can reliably perform the given task. For some tasks, what's required for reliability is relatively easy to specify. In the context of 18th century tests of theories of gravity, for instance, (1) is reliable if and only if it delivers appropriately accurate answers for the value of acceleration given accurate measurements of string length for a low-amplitude seconds pendulum. For other tasks, it's much more difficult to say exactly what reliability entails. Where the model generates a representation of an earth-like climate that's then used to determine the "forcing" effects of increasing atmospheric CO_2 concentration (see Winsberg 2018, 64-69), it's hard to state more precise conditions on reliability than that the representation is *appropriately* earth-like—that it captures all of the features our system that are relevant to the forcing question. Differences between tasks engender differences in what's required for the auxiliary hypothesis to reliably perform the task. One implication that we've already encountered is that it's misleading to identify this question with the question of whether the auxiliary hypothesis can serve as a premise in a deduction.

The context-sensitivity of reliability is central to explaining the phenomenon that we encountered in the counterexamples: in both of the cases considered above, the model is known to be capable of performing the tasks that it is given even though it is false or relies on false assumptions. With (1), physicists knew that it reliable enough for converting values of observable features to values for acceleration so long as the amplitude was low. Similarly for the HW model: population geneticists know that the HW model can reliably isolate the effect of heterozygote advantage so long as the population is large enough. In order to perform this function, the model doesn't need to have true assumptions. In fact, it had better not: if there's a real effect to isolated, then the model must have some (specific) false assumptions. Reliably performing the function of isolating an effect or converting one value to another doesn't depend on the model being true; we only require the latter when we're using the auxiliary hypotheses as premises. The arguments of the last two sections were aimed at showing that auxiliary hypotheses aren't primarily used as premises in arguments, however. Their roles are more specialized.

4.2 The implications for holism

The second aspect of the epistemic tool view is a transformation of holism. Consider a thermometer that consistently gives a temperature value that is 20° F above the actual value. Taken independently of its context, we'd be inclined to say that this thermometer isn't reliable (in the sense that we're interested in); it doesn't deliver accurate answers to questions about the current temperature. But a thermometer of this sort isn't useless: if we know how the instrument behaves, we can correct for the discrepancy between the reading and the real temperature (compare S. D. Norton & Suppe 2001). This "correction" can go the other way as well, as we saw in the pendulum case. Since we know that the equation is only reliable at low amplitudes, we can constrain our physical pendulum to ensure that these conditions are met. And thus, much as the Duhemian finds holism in the fact that multiple premises must be tested together, the epistemic tool theorist must acknowledge in any given case, we're testing not just the hypothesis of interest but also the reliability of various different tools.

The resulting holism differs in one crucial respect from the holist picture encapsulated in Duhem's (P3). The Duhemian presents a picture in which the sentences of a theory are tied together by a series of logical relationships that render the various commitments inseparable in principle. On the epistemic tool view, by contrast, the relationship between different epistemic tools employed in a test isn't *logical*; it's not a matter of entailments between propositions. Instead, it's a matter of functionality: there's holism in the sense of relationships of interreliance between tools that can only properly function in the presence of other, properly-functioning, tools. Importantly, because the resulting holism is matter of function rather than a logical relationship, the degree of separability can vary between cases. In some cases, we're able to isolate and confirm the reliability of individual tools; in others, we're only be able to evaluate the reliability of a large group of commitments together. The degree of separability here depends on our ability to identify either (a) other cases in which a particular epistemic tool can be employed to perform the same task in a different context or (b) other auxiliary hypotheses capable of performing the same task. In the first case, we can use these other contexts to evaluate the reliability of the tool for the task; in the second case, we can use robustness reasoning to check the reliability of the particular auxiliary hypothesis for the task and to lessen our reliance on the reliability of that tool.

To see why there is a difference between holism as instantiated in the logical relationships between sentences and the holism defended here, an analogy will be helpful. Consider the following two inductive arguments:

(2) All observed swans are white, therefore the next swan will be white.

(3) All observed swans are white, therefore all swans are white.

For me, at least, the premise of both arguments is true: I have never seen a black swan. And I take it that I have quite good reason to believe (2). After all, I expect to pass a pond on my way home this afternoon, and there are often white swans there. My present and nearby contexts are very similar to my past contexts. This similarity provides me with good reason to expect generalization to be reliable in the near future. By contrast, I don't have good reason to believe (3)—in fact, I have good reason to believe that it is false. But even if I lacked such reasons, I take it that the minimal background theory required to justify (3) is much stronger than the minimal background theory required to justify (2).[8]

There is a strong analogy between (a) assuming that (1) is reliable and assuming that it is true on the one hand and (b) the inference to "the next swan is white" and the inference to "all swans are white" on the other. (The force of the analogy can be strengthened by recognizing that the latter contrast is between the assumption that "all swans are white" is a reliable generalization in the present context and assuming that it is true.) In both cases, we require (substantially) weaker background assumptions to justify treating

[8] I'm implicitly assuming a "materialist" account of induction here (see J. Norton 2003), but I don't think that this claim relies on any controversial aspect of Norton's account.

generalizations based on past experience as reliable in local circumstances than we do to justify treating them as true. In order to show that an equation like (1) is true (or that the HW model is "perfect" in the terms of Teller 2001), we'd need to demonstrate that, for example, (1) holds in low-gravity areas such as the moon and in cases where the force is something other than gravity. We would need, in other words, an entire theory of pendular motion that would account for all the different contexts that the pendulum could be in and address how it would behave in each context. In order to demonstrate that it is reliable in the current earth-bound, low-amplitude context, by contrast, we don't need to take a stance on what would happen in those other cases. We don't need a "whole theory" of pendular motion. We need some relatively low-level empirical results—what Azzouni (2000) calls "gross empirical regularities"—and the ability to extend them marginally to cases that are very similar to the ones previously encountered. Nothing like Duhem's (P3) can be defended in the context of the epistemic tool view, therefore: at most we're committed to the reliability of some limited aspects of a given theory in a particular context, but this instrumental reliance on a part cannot be transformed into a reliance on the truth of the whole.

4.3 The role of physical and theoretical instruments

The final implication of the view is that the reliability of physical and theoretical instruments ought to play the same role in our confidence in a result. Suppose that the thermometer reads 18° F when I look out my window in the morning, for example. In a Bayesian framework, my confidence that it is in fact freezing should depend on how reliable I believe the thermometer to be. This can be seen by looking at Bayes' theorem:

$$P(h|d) = \frac{P(h)P(d|h)}{P(h)P(d|h) + P(\neg h)P(d|\neg h)} \tag{4}$$

The unreliability of the thermometer will figure into the value of $P(d|\neg h)$. It can also be seen (more directly) by noting that if we introduce a variable, r, expressing the reliability of the thermometer, the posterior probability can be expressed as

$$\begin{aligned} P(h|d) &= P(r)P(h|d,r) + P(\neg r)P(h|d,\neg r) \quad (5) \\ &= P(r)P(h|d,r) + P(\neg r)P(h) \quad (6) \end{aligned}$$

After all, if my thermometer has simply gone haywire, I have no more reason to think it will deliver false positive than false negatives, and so should simply treat the probability of h given $\neg r$ as equivalent to the prior—a stipulation that Bovens & Hartmann (2003, 57) term "Edman's condition" after Edman (1973); see also Mayo (2018, 5).[9]

Once we recognize that the models or theories employed in a simulation are akin to the tools of a more traditional experiment, it becomes clear that the same reasoning applies in the modeling case: our confidence in h should depend on our confidence that the model is performing the way it should. This is true regardless of whether we evaluate this quantity indirectly, as part of what goes into $P(d|\neg h)$ or build it into our calculations explicitly using a reliability variable. But notice that it must be a reliability variable, and not something like $P(m)$ where this is interpreted as the probability that the model is true. After all, if we know that the model is idealized, we know that $P(m)$ is zero (if it isn't a category mistake). The more traditional view can't treat models and instruments the same in this manner—or at least, can't make sense of why we should—while the epistemic tool view can. As recent work on simulations (e.g. Mäki 2005; Parker 2009b; Winsberg 2010, chapter 4) has emphasized, it's virtually impossible to draw epistemically interesting general distinctions between tests that employ physical instruments and those that employ theoretical ones; the distinction to draw is the more case-by-case one of whether the tool in question is appropriate for the task. The epistemic tool view makes it clear why this should be the case: both physical instruments and theoretical ones contribute to the reliability of a given test, and in the same way.

5 Conclusion

The primary lesson of Austin's *How to do Things With Words* is that words are used for far more than assertion. The main lesson of this paper runs parallel: theory—in the broad sense—is used for far more than (truth-apt) representation. More specifically, I've argued against a traditional Duhemian account of auxiliary hypotheses, a view in which background theories and models are primarily employed as premises in an argument. This Duhemian view is mistaken because background theories and models are not always (perhaps not usually) asserted; that is, they're not used to truthfully represent. Instead, they're used to translate, isolate, identify, amplify, control, or correct. And just as the conditions on when one is correct in asserting a sentence differ from those under which one is correct in employing the same content as a question, the conditions under which it is correct to use theory for one of these tasks differ from those under which it is correct to say that the theory is true. On the resulting view, the acceptability of an auxiliary hypothesis depends not on its truth but on contextual factors such as the task or purpose it is put to and the other tools employed alongside it, and this result has implications for our understanding of holism and the relationship between physical and theoretical tools.

[9] Of course, Edman's condition is an idealization. We might expect an unreliable thermometer (like an unreliable clock) to be relatively close. *How* an instrument is broken matters, and we'll rarely have *no* information on that front. The same point can be made about theoretical instruments, however, and so the idealization shouldn't undermine the point.

Throughout the argument for this conclusion, I've largely avoided discussion of either the various anti-realist conclusions that are typically drawn from the holist argument or the proper historical interpretation of Duhem himself. On the former front, I'm not sure I have much to add to the literature that this essay draws on—in a sense, I'm working backwards from the critiques of (across-the-board) holism and underdetermination offered by Azzouni (2000), Hacking (1983), and Longino (1990) to the implications for testing and auxiliary hypotheses. I do have one suggestion on the latter front, however. Holism is usually seen as one of the major motivators for anti-realism of various sorts, and it would be natural to assume that Duhem's own version of anti-realism follows from his commitment to holism. But this is not how *Aim and Structure* is written: the anti-realism appears first; arguments for holism and underdetermination are given much later. The argument that I have offered in this paper relies—quite explicitly—on drawing a sharp distinction between the (in-context) reliability of a theory and its truth, a distinction that is at the very least complicated by Duhem's rejection of the idea that science provides us (or aims to provide us) with true theories at all. It is worth suggesting, therefore, that Duhem himself is a holist not because of a mistaken identification of the use a theory for testing and using it for derivation, but because of a purposeful one motivated by his earlier commitment to a form of anti-realism about the aims or prospects of science—a commitment he seems to derive not from underdetermination arguments but rather from his study of the history of science (Brenner 1990, 332-33). I do not believe that this sort of anti-realism alone would rescue the argument, but perhaps it would make better sense of Duhem's formulation of the argument itself.

References

Ariew, R. (2018). Pierre Duhem. *Stanford Encyclopedia of Philosophy.* URL: https://plato.stanford.edu/archives/fall2018/entries/duhem/.

Austin, J. L. (1962). *How to Do Things With Words.* Cambridge, MA: Harvard University Press.

Azzouni, J. (2000). *Knowledge and Reference in Empirical Science.* London: Routledge.

Bokulich, A. (forthcoming). Using Models to Correct Data: Paleodiversity and the Fossil Record. *Synthese.*

Bovens, L. & S. Hartmann (2003). *Bayesian Epistemology.* Oxford: Oxford University Press.

Brenner, A. A. (1990). Holism a Century Ago: The Elaboration of Duhem's Thesis. *Synthese* 83.3: 325–35.

Cartwright, N. (1991). Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy* 23.1: 143–55.

Currie, A. (2018). *Rock, Bone, and Ruin: A Optimist's Guide to the Historical Sciences.* Cambridge, MA: MIT Press.

Duhem, P. (1914/1951). *The Aim and Structure of Physical Theory*. Trans. by P. P. Wiener. Princeton, NJ: Princeton University Press.

Edman, M. (1973). Adding Independent Pieces of Evidence. In: *Modality, Morality and Other Problems of Sense and Nonsense*. Ed. by S. Halld/'en. Lund: Gleerup: 180–8.

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

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

Filion, N. & R. H. C. Moir (2018). Explanation and Abstraction From a Backward-Error Analytic Perspective. *European Journal for the Philosophy of Science* 8: 735–59.

Hacking, I. (1983). *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

Houkes, W. & K. Vaesen (2012). Robust! Handle with Care. *Philosophy of Science* 79.3: 345–64.

Lewontin, R. C. (1974). *The Genetic Basis of Evolutionary Change*. New York, NY: Columbia University Press.

Lloyd, E. (2012). The Role of "Complex" Empiricism in the Debates about Satellite Data and Climate Models. *Studies in History and Philosophy of Science Part A* 43.2: 390–401.

Longino, H. (1990). *Science and Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Mäki, U. (2005). Models are Experiments, Experiments are Models. *Journal of Economic Methodology* 12.2: 303–15.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.

— (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Morgan, M. S. & M. Morrison (1999). Models as Mediating Instruments. In: *Models as Mediators: Perspectives on Natural and Social Science*. Ed. by M. S. Morgan & M. Morrison. Cambridge: Cambridge University Press: 10–37.

Morrison, J. (2010). Just How Controversial is Evidential Holism? *Synthese* 173.3: 335–52.

Morrison, M. (2015). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford: Oxford University Press.

Norton, J. (2003). A Material Theory of Induction. *Philosophy of Science* 70.4: 647–70.

Norton, S. D. & F. Suppe (2001). Why Atmospheric Modeling Is Good Science. In: *Changing the Atmosphere: Expert Knowledge and Environmental Governance*. Ed. by C. A. Miller & P. N. Edwards. Cambridge, MA: The MIT Press: 67–105.

Odenbaugh, J. & A. Alexandrova (2011). Buyer Beware: Robustness Analyses in Economics and Biology. *Biology & Philosophy* 26.5: 757–71.

Orzack, S. & E. Sober (1993). A Critical Assessment of Levins's "The Strategy of Model Building in Population Biology" (1966). *The Quarterly Review of Biology* 68.4: 533–46.

Parker, W. S. (2009a). Confirmation and Adequacy-for-Purpose in Climate Modeling. *Proceedings of the Aristotelian Society* 83: 233–49.

— (2009b). Does Matter Really Matter? Computer Simulations, Experiments, and Materiality. *Synthese* 169.3: 483–96.

— (2016). Reanalyses and Observations: What's the Difference? *Bulletin of the American Meteorological Society* 97.9: 1565–72.

Quinn, P. (1974). What Duhem Really Meant. In: *Methodological and Historical Essays in the Natural and Social Sciences*. Ed. by R. S. Cohen & M. W. Wartofsky. Dordrecht: Reidel: 33–56.

Schliesser, E. & G. E. Smith (forthcoming). Huygens's 1688 Report to the Directors of the Dutch East India Company on the Measurement of Longitude at Sea and the Evidence it Offered Against Universal Gravity. *Archive for the History of the Exact Sciences*.

Sober, E. (1984). *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. Chicago: University of Chicago Press.

— (1999). Testability. *Proceedings and Addresses of the American Philosophical Association* 38.2: 47–76.

Strevens, M. (2001). The Bayesian Treatment of Auxiliary Hypotheses. *British Journal for the Philosophy of Science* 52: 515–37.

Suppes, P. (1962). Models of Data. In: *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*. Ed. by E. Nagel, P. Suppes, & A. Tarski. Stanford: Stanford University Press: 252–61.

Teller, P. (2001). Twilight of the Perfect Model Model. *Erkenntnis* 55.4: 393–415.

Tempelton, A. R. (2006). *Population Genetics and Microevolutionary Theory*. Hoboken, NJ: John Wiley & Sons.

Wilson, M. (2006). *Wandering Significance: An Essay on Conceptual Behavior*. Oxford: Oxford University Press.

Winsberg, E. (2010). *Science in the Age of Computer Simulation*. Chicago, IL: University of Chicago Press.

— (2018). *Philosophy and Climate Science*. Cambridge: Cambridge University Press.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

— (2006). Some Varieties of Robustness. *Journal of Economic Methodology* 13.2: 219–40.