What's Wrong With Our Theories of Evidence?[*]

Julian REISS

Received: 18.01.2014
Final version: 29.03.2014

BIBLID [0495-4548 (2014) 29: 80; pp. 283-306]
DOI: 10.1387/theoria.10782

ABSTRACT: This paper surveys and critically assesses existing theories of evidence with respect to four desiderata. A good theory of evidence should be both a theory of evidential support (i.e., be informative about what kinds of facts speak in favour of a hypothesis), and of warrant (i.e., be informative about how strongly a given set of facts speaks in favour of the hypothesis), it should apply to the non-ideal cases in which scientists typically find themselves, and it should be 'descriptively adequate', i.e., able to adequately represent typical episodes of evidentiary reasoning. The theories surveyed here—Bayesianism, hypothetico-deductivism, satisfaction theories, error statistics as well as Achinstein's and Cartwright's theories—are all found wanting in important respects. I finally argue that a deficiency all these theories have in common is a neglect or underplaying of the epistemic context in which the episode of evidentiary reasoning takes place.

Keywords: Scientific method; evidence; Bayesian confirmation theory; error statistics; contextualism.

RESUMEN: Este artículo describe y valora críticamente diversas teorías de la evidencia en relación a cuatro desiderata. Una buena teoría de la evidencia debería ser tanto una teoría sobre el apoyo evidencial [*evidential support*] (ser informativa sobre qué tipos de hechos hablan a favor de la hipótesis) como sobre la justificación [*warrant*]; debería aplicarse en las situaciones no ideales en las que normalmente se encuentran los científicos; y debería ser 'descriptivamente adecuada', esto es, capaz de representar correctamente episodios típicos de razonamiento evidencial. Las teorías aquí revisadas—bayesianismo, hipotético-deductivismo, teorías de la satisfacibilidad, la estadística del error, así como las propuestas de Achinstein y Cartwright—se consideran deficientes en aspectos básicos. Argumentaré que un defecto común en todas ellas es que olvidan, o minusvaloran, el contexto epistémico en el que el episodio de razonamiento evidencial tiene lugar.

Palabras clave: método científico; evidencia; teoría bayesiana de la confirmación; estadística del error; contextualismo.

1. Introduction

Peter Achinstein once argued that 'Philosophical theories of evidence are (and ought to be) ignored by scientists' (Achinstein 1995, title and passim). His reasons were that 'such theories propose concepts of evidence that (a) are too weak to give scientists what they want from evidence and (b) make the evidential relationship a priori, whereas typically establishing whether e if true is evidence that h requires an empirical investigation'. I wholeheartedly agree with (b), and will argue as much below. (a) is, however, too strong a claim. It is certainly not the case that all existing philosophical theories of evidence propose concepts of evidence that are too weak. Some are also too strong, and most are both too strong and too weak at the same time. The issue with these theories is not that they propose concepts that are too weak but that they propose the

[*] Financial support from the project FFI2012-37354/CONSOLIDER INGENIO CSD2009-0056 of the Spanish Government is gratefully acknowledged.

wrong concepts: philosophical concepts that have little to do with the needs of practising scientists.

The main aim of this paper is to survey and critically examine existing philosophical theories of evidence. It will turn out that most accounts don't even get commonplaces such as 'correlations are (or provide) evidence for causal relations' right. While the main aim is make a critical inventory of what there is, the paper will also suggest a way for improvement and thus lay the ground for a more satisfactory theory.

2. Desiderata

Before diving into the survey of existing theories of evidence let us lay out a number of desiderata any such theory should fulfil in order to have a standard against which to judge. To formulate our desiderata it is necessary to first distinguish two different concepts of evidence.

When we say we have evidence *e* for a scientific hypothesis *h*, we may have either of two importantly different meanings in mind (Salmon 1975). We might either mean that *e* is a 'mark' or 'sign' or 'symptom' of the hypothesis' being true, that *e* is a '*piece* of evidence' for *h*. The suspect's fingerprints on the murder weapon, another suspect's incontrovertible alibi, the appearance of a shirt splattered with what turns out to be the victim's blood in the suspect's laundry basket, an insurance police demonstrating that the suspect profits a great deal from the victim's death each speak in favour of the hypothesis that the suspect murdered the victim without constituting (yet) *a good reason* to infer it. This notion of evidence has therefore also been referred to as 'supporting evidence' (Rescher 1958, 83). I will, more concisely, call it 'support'.

Alternatively, when we say that we have evidence *e* for a scientific hypothesis *h*, we may mean that we have a '*body* of evidence' for *h* that constitutes 'proof' that *h*, that licenses *h* or that *e* constitutes a '(weak, strong *etc.*) reason to infer' *h*. All the pieces of evidence just mentioned together constitute a body of evidence which in the right circumstances may demonstrate that the suspect must be guilty or at least warrant the hypothesis to a certain degree. The latter kind of evidence I will call therefore 'warranting evidence' or short 'warrant'.

We cannot have warrant for a hypothesis without having support. We cannot have a licence to infer a hypothesis without being in the possession of facts that speak in its favour. Conversely, it is at least conceivable, and in my view frequently true, that we are in the possession of a fact that speaks in favour of the hypothesis without having any reason whatsoever to infer the hypothesis. Suppose the second suspect's fingerprints are also on the murder weapon. This is no doubt a fact that speaks in favour of the hypothesis that the second suspect committed the crime. However, given that (let us suppose) there is a good alternative account for the presence of fingerprints and an incontrovertible alibi, the hypothesis is not warranted, not even minimally. To give another example, a patient's cough supports the hypothesis that the patient suffers from lung cancer but also countless other conditions. The support relation remains in place even when there are other facts that (let us suppose) conclusively rule out the lung cancer hypothesis. Support is analytically more basic than warrant: we need the former to have the latter but we don't need the latter to have the former.

Support and warrant are related but not identical to the Bayesian notions of incremental and absolute confirmation. According to the Bayesian theory, e incrementally confirms h if and only if $\text{Prob}(h \mid e) > \text{Prob}(h)$; e absolutely confirms h if and only if $\text{Prob}(h \mid e)$ is high. A brief discussion of the base-rate fallacy shows that support is like incremental confirmation and warrant like absolute confirmation in some cases. Suppose that a patient is worried that he might suffer from a disease *d*. There is a really good test available, which has a false negative rate of zero and a false positive rate of 5%. The patient takes the test and gets a positive result (e = positive). What is the probability that the patient has the disease given he has the positive result? Many people fallaciously argue 95% because they mistake the posterior probability $\text{Prob}(h \mid e)$ with the 1 – the false positive rate $\text{Prob}(e \mid \neg h)$. In fact the posterior is much lower. Suppose only 1 in 1000 people carry the disease: $\text{Prob}(h = \text{patient has } d) = 0.1\%$. The false positive and negative rates give us the likelihoods: $\text{Prob}(e \mid h) = 1$ and $\text{Prob}(e \mid \neg h) = 5\%$. The posterior probability is given by Bayes' Theorem:

$$\text{Prob}(h \mid e) = \text{Prob}(h)^{*}\text{Prob}(e \mid h)/[\text{Prob}(e \mid h)^{*}\text{Prob}(h)+\text{Prob}(e \mid \neg h)^{*}\text{Prob}(\neg h)]$$
$$= 0.001^{*}1.00/[1.00^{*}.001+.05^{*}.999] = .019627 \approx 2\%$$

The positive test result incrementally confirms the hypotheses $\text{Prob}(h \mid e) \approx 2\% > .1\% = \text{Prob}(h)$. But it does not absolutely confirm the hypothesis as 2% is not 'high'. Similarly, a positive test result certainly speaks in favour of the hypothesis and therefore supports it but, given these numbers, does not constitute a good reason to infer it.

More generally, under a sensible assignment of probability functions, (a suitable description of) most facts that support a hypothesis will also raise its probability: (an evidential statement describing) fingerprints on the murder weapon will raise the probability of the hypothesis that whoever left the fingerprints was the murderer, (an evidential statement describing) symptoms will raise the probability of the hypothesis that the disease is present and so on. Probability raising is, however, at best necessary but not sufficient for support: a random individual's presence near the crime scene will raise the probability that she committed the crime but it does not support the hypothesis; the suspect's murdering the victim will raise the probability that she killed him but not support it in the sense of being a mark or symptom for its truth (more on this below).

Further, while warrant is a notion that admits of degrees: a hypothesis can be weakly or strongly warranted, there can be more or less good reasons to infer a hypothesis, there is no reason to suppose that warrant is always quantitative (by which I mean measurable on at least a cardinal scale) and can be represented by probabilities.[1] The 'high probability' theory of warrant also has some technical problems that I will discuss below. The concepts of incremental confirmation and support on the one

[1] The idea is that we might be able to *rank* hypotheses with respect to their degree warrant without the numbers we assign to their ranks being meaningful beyond representing their place in the ranking. Thus, if we assign (say) zero to 'no warrant', one to 'proof', 0.4 to 'weak warrant' and 0.8 to 'strong warrant', we only know that the strongly warranted hypothesis is more highly warranted than the weakly warranted hypothesis; we do not know that it is *twice* as strongly warranted.

hand and of absolute confirmation and warrant on the other overlap, but they are not identical.[2]

A good theory of evidence should be a theory of both support and warrant. One way to motivate this is by invoking Carnap's *Principle of Total Evidence*. Carnap writes (Carnap 1947, 138-9):

> A principle which seems generally recognized, [footnote omitted] although not always obeyed, says that if we wish to apply such a theorem of the theory of probability [a theorem which states the degree of confirmation of a hypothesis with respect to the evidence] to a given knowledge situation, then we have to take as evidence e the total evidence available to the person in question at the time in question, that is to say, his total knowledge of the results of his observations. [footnote omitted]

This way formulated, the principle would hardly be practicable. There is no way to take all known facts into account when evaluating a hypothesis, neither for an individual researcher nor for a scientific community. Thus, Carnap quickly appends the principle by stating that *irrelevant* additional items of evidence may be omitted (139). So we need criteria of evidentiary relevance: criteria that tell us what kinds of facts we have to collect in order to assess a given hypothesis. These criteria are delivered by a theory of support.[3]

But we can't only have criteria of relevance. We also need criteria to tell us how to assess the hypothesis, given the facts we've collected in its support; or, conversely, criteria that tell us how much support of what kind we need in order to achieve a given degree of warrant. In other words, what we require is criteria that translate between knowledge of the facts relevant to assessment of a hypothesis and judgements about the hypothesis.

A theory of evidence that didn't tell us about support would be impracticable; a theory that didn't tell us about warrant would not be useful. Here, then, are our first two desiderata for a good theory of evidence: it should be a theory of both support *and* of warrant.

Our third desideratum is that the theory applies to non-ideal cases. Practising scientists often (if not always) have to gather evidence and assess hypotheses in situations where perfectly controlled experiments and randomised trials are unavailable (for whatever reason: technological, financial, ethical) and background knowledge is scarce or lacking in reliability. A theory, say, that regarded as evidence exclusively that which was produced by a flawless randomised trial would presumably get it right when it applies, but it would hardly ever apply and therefore not be of much use. A practicable theory must be able count as evidence that which has been produced under the condi-

[2] Similar remarks apply to the 'qualitative'/'quantitative' confirmation pair. 'Support' means 'positive relevance' or 'speaking in favour' and is a qualitative concept. However, this does not mean that support necessarily has to be cashed out in terms of standard qualitative confirmation theories such as the hypothetico-deductive theory or satisfaction theories such as Hempel's or Glymour's. See below for a criticism of these theories. I already discussed that warrant comes in degrees but isn't necessarily quantitative.

[3] A fact that speaks against a hypothesis is relevant to its assessment but does not support it. However, a fact that speaks *against* hypothesis *h* will speak *in favour of* an alternative hypothesis *h′*. We therefore focus on support and ignore *in*firming evidence here.

tions in which typical scientists find themselves and deliver verdicts on evidence of this kind.

Lastly, our theory of evidence should be 'descriptively adequate'. That is, it should, by and large, regard as evidence what practising scientists regard as evidence and confer assessments on hypotheses roughly in line with practice. It is clear that most theories, whether in science or philosophy or elsewhere, are idealised accounts of the facts. Moreover, if a theory is to have normative import it can deviate from practice to the extent that practice errs (in that, say, it regards as evidence what shouldn't be so regarded or delivers bad verdicts about hypotheses). But these deviations must be *excusable*. For each deviation it must be possible to tell a story why it is either harmless or beneficial (because improving on practice) or something similar.

In sum, a theory of evidence should:

- be both a theory of support as well as
- a theory of warrant;
- apply to non-ideal scenarios; and
- be descriptively adequate.

3. Existing Theories of Evidence

In this section I will examine how well existing accounts fare with respect to the four desiderata I laid out above. Throughout, I will assume that a correlation between two variables *I* and *D*[4] (where, say, *I* = smoking or *I* = money and *D* = lung cancer incidence or *D* = prices) is evidence for the hypothesis '*I* causes *D*', which should be uncontroversial.

3.1 Bayesianism

According to the Bayesian theory,

(**BAY**) *e* is evidence for *h* if and only if $\text{Prob}(h \mid e) > \text{Prob}(h)$,

where *e* is an evidential statement, *h* is the scientific hypothesis, Prob(*h*) denotes the probability that *h* and $\text{Prob}(h \mid e)$ the probability that *h*, given *e*.

Bayesianism appears to give us a criterion of relevance: to support a hypothesis *h*, collect all and only those facts learning about which raises the posterior probability of *h*. This characterisation makes clear, however, that Bayesianism puts the cart before the horse if we understand it as a theory of support. The instruction 'collect all and only those facts learning about which raises the posterior probability of *h*' is not one that allows us to identify which facts we have to look for to begin with. Facts don't come with probabilities attached. Perhaps, once we have accepted a new fact as evidence for a hypothesis, we will raise its posterior. But we have to know if the fact is evidence before we can decide how to adjust our probabilities.

[4] I call the variables *I* for independent and *D* for dependent variable instead of, say, *C* and *E* for cause and effect in order to indicate that the causal relation is merely putative.

When a scientists learns a new evidential statement (say, *e* = 'The result of the experiment is *r*'), she must already know whether or not *e* is relevant to her hypothesis and whether it speaks in favour of or against the hypothesis in order to know whether her posterior of *h* on *e* is higher or lower than or equal to her prior. Change in degree of belief is thus an epiphenomenon of evidential relevance (Glymour 1980). The Bayesian machinery allows her to form and revise her beliefs rationally, but it silent on the question of which beliefs are evidence in the first place. Colin Howson and Peter Urbach are thus entirely correct in saying (Howson and Urbach 1993, 272):

> The Bayesian theory we are proposing is a theory of inference from data; we say nothing about whether it is correct to accept the data [...] The Bayesian theory of support is a theory of how the acceptance as true of some evidential statement affects your belief in some hypothesis.

If Howson and Urbach are right, Bayesianism is at best a theory of warrant, not one of support. Unfortunately, it frequently gets its assessments of warrant wrong. Suppose that background knowledge tells us that learning that *I* and *D* are correlated (*e*) raises the probability that *I* causes *D* (*h*) and thus constitutes evidence according to the Bayesian theory. The problem is that the correlation between *I* and *D* is at best a *sign of the truth* of the hypothesis, not in itself *a good reason to infer it.* Understood as a theory of warrant, Bayesianism is thus too weak, just as Achinstein argues (*cf.* Achinstein 2001).

One possible way to fix it is to posit a threshold level *x* such that *e* is warranting evidence for *h* iff Prob(*h*) ≤ *x* and Prob(*h* | *e*) > *x*. But that will fail, no matter what one chooses as one's *x*. Take a salient choice, *x* = .5. Suppose one knows nothing about whether or not *I* causes *D*. Objective Bayesians would therefore, say, on the principle of indifference, assign Prob(*h*) = Prob(¬*h*) = .5. In this case, of course, anything that supports the hypothesis also constitutes a good reason to infer it. But this would be a mistake.

So let's raise *x* to, say, .6. Now everything depends on the likelihoods. Suppose then that it is very likely that if *I* causes *D*, *I* will be correlated with *D*: Prob(*e* | *h*) = .9, and that if *I* does not cause *D*, the correlation is rather unlikely: Prob(*e* | ¬*h*) = .3. In this case, Prob(*h* | *e*) = 3/4. Given that the probabilities the Bayesian requires are not empirically ascertainable, it will always be possible to rig the numbers in such a way that the theory yields the 'wrong' result, as it does here.

Further, the Bayesian theory yields the wrong result for necessary conditions for a hypothesis' being true which do not support it. If there wasn't any aflatoxin, the substance could not cause liver cancer. Thus, Prob(aflatoxin causes liver cancer | aflatoxin exists) > Prob(aflatoxin causes liver cancer | aflatoxin does not exist) = 0. No biomedical scientist would regard the existence of aflatoxin as supporting the hypothesis that it causes cancer. (For one thing, the existence of aflatoxin does not speak in favour of *h as opposed to* ¬*h*.)

There is also the reverse case but that may be less of a problem. Take two statements *h*: Aflatoxin is a carcinogen' and *e*: 'Aflatoxin is a potent carcinogen'. Since *e* entails *h*, Prob(*h* | *e*) > Prob(*h*) unless the latter is already unity, and Prob(*h* | *e*) > *x* for any choice of *x* < 1. Scientists would probably be loath to call statements such as *e* evidence for *h* in either of our two senses. But arguably, the fact that aflatoxin is a *potent* carcinogen constitutes a good reason for inferring that aflatoxin is a carcinogen—perhaps the best reason there is.

Bayesianism has no problems dealing with non-ideal scenarios, as long as background knowledge dictates the right probabilities. The following set of probabilities is perfectly plausible under a Bayesian account of evidence: Prob(*h* = 'aflatoxin is carcinogenic') = Prob(¬*h*) = .5; Prob(*h* | *e* = 'aflatoxin exposure and liver cancer incidence are correlated', b_1 = '*e* was recorded in an observational study that is potentially subject to bias') = .6 and Prob(*h* | *e*, b_2 = '*e* was recorded in a well-designed randomised trial') = .9, so that a correlation is evidence for *h* in both scenarios but stronger evidence when it was recorded in an experiment than when it was recorded in an observational study.

There is much debate in the literature about whether or not Bayesianism constitutes an adequate model for scientific reasoning (to mention but three contributions, see Howson and Urbach 1993 for a defence of subjective Bayesianism, Williamson 2010 for a defence of objective Bayesianism and Mayo 1996 for an anti-Bayesian account of scientific reasoning). Let me discuss just one oddity here. If we know nothing at all about hypotheses *h* = 'aflatoxin is carcinogenic' and ¬*h* = 'aflatoxin is not carcinogenic', the objective Bayesian will assign them both a weight of .5. (The subjective Bayesian can assign any weight at all but the numbers do not matter to my point.) So complete ignorance is one way to come up with the judgement that Prob(*h*) = .5. An alternative route is, for example, this. Aflatoxin belongs to a class of substances, most of which are not carcinogenic so that (say) the prior probability given only that background knowledge is .1. But then evidence comes in: aflatoxin exposure and liver cancer incidence are correlated in humans; experiments with animal models show that exposure causes liver cancer in at least some species; while all observational studies (related to this hypothesis) may well be confounded, it has been shown that at least some of the possible causes of liver cancer that are endemic in populations exposed to aflatoxin cannot explain incidence rates; etc. On the other hand, many animal models appear to be resistant to aflatoxin and there is no reason to think that those models that are susceptible are better models for humans than those that are not; and there remain numerous confounders. Suppose that at the end of the day on the balance of evidence Prob(*h*) = Prob(¬*h*) = .5. Intuitively, this latter situation where a great deal of evidence leads us to assign the probabilities seems to be very different from the situation where they stem from ignorance. Bayesianism, however, has no resources to distinguish between the two situations. Scientists do regularly distinguish between having positive grounds for thinking that a hypothesis has a chance of being true that is strictly between zero and one and not knowing at all. This speaks against Bayesianism's descriptive adequacy.[5]

[5] (Norton 2011) contains a far more detailed discussion of the difficulties Bayesianism has to represent ignorance.

3.2 Hypothetico-deductivism

According to the hypothetico-deductivist theory,

(**HD**) e is evidence for h if and only if h deductively entails e.

Unlike Bayesianism, (**HD**) tells us what facts to watch out for: those facts a description of which is entailed by the hypothesis. Unfortunately, deductive entailment is neither necessary nor sufficient for support. Causal relations *typically* issue in (probabilistic) regularities but not always. Any given causal hypothesis of the form 'I causes D' does not, therefore, entail the corresponding regularity or probabilistic claim such as 'Whenever I, then D' or '$\text{Prob}(I \mid D) > \text{Prob}(I)$'. Nevertheless, regularities and claims about probability raising (typically) support causal claims. Causal hypotheses do, however, entail existential claims, as we've seen above. And as we've also seen, the existential claim is not relevant to the truth of a causal hypothesis. Hypothetico-deductivism is therefore not a good theory of support.

Hypothetico-deductivism does not distinguish between support and warrant. Suppose that we are in the rare position to have background knowledge b that is strong enough such that given this background knowledge, the causal hypothesis h 'I causes D' does entail the statement e 'I and D are correlated'.[6] Now e comes out correctly as speaking in favour of h but—read as a theory of warrant, wrongly—as presenting a good reason to infer h. e is not a good reason to infer h because it is possible that e is also entailed by alternative hypotheses h', h'', h''' and so on that are incompatible with h (in conjunction with b). Unless the alternatives have been ruled out, e might speak in favour of h but it does not speak in favour of h as opposed to h', h'', h''' and so on. We therefore have no good reason to choose h over any of its alternatives.

As it is silent on what conditions the implications of h should be observed, it applies equally to ideal and non-ideal situations.[7] Whether the theory is descriptively adequate depends on how it is interpreted. Few scientists will subscribe to a strict reading of the theory according to which a correlation is not relevant to the assessment of a hypothesis. On a looser reading of the deductive part of the theory according to which the evidence e is given by what a scientist can expect to observe if the hypothesis were true, it may well constitute an accurate description of scientific practice (Reiss forthcoming a).

3.3 Satisfaction Theories

Hempel's 1945 satisfaction account of evidence is a development of the idea that an instance of a generalisation is evidence for the generalisation. As its main problems are

[6] In her 2001 Nancy Cartwright lists the conditions that have to be in place so that causal relations will issue in correlations. My own view is that there are indefinitely many reasons for which I and D can fail to be correlated despite the fact that I causes D. So strictly speaking, there is no amount of background knowledge such that $h\&b \vdash e$.

[7] Or rather, as few scientific hypotheses entail any statement about observations on their own, the hypothesis h has to be conjoined with background knowledge b in order to make predictions. But since there are no constraints on admissible k's, one is free to formulate b in such a way as to describe ideal as well as non-ideal situations of evidence gathering.

well known—the ravens paradox, the grue problem, the problem that statements cast in an observational language cannot constitute evidence for hypotheses cast in a theoretical language (see Norton 2010 for a detailed discussion)—I will not consider the theory it in any detail here. One important issue is that generic causal claims such as 'Aflatoxin causes liver cancer (in humans)' do not appear to be straightforward generalisations of causal claims concerning individuals such as 'Wei's exposure to aflatoxin caused his liver cancer'. The generic claim can neither be analysed as 'For all *i*, *i*'s exposure to aflatoxin causes *i* to develop liver cancer' (because only a minority of those exposed will develop the disease) nor as 'There exists an *i* such that *i*'s exposure to aflatoxin causes *i* to develop liver cancer' (because this is too weak: some people die in car accidents because they wore a seat belts; but seat belts save lives and don't cause deaths, see Hitchcock 1995; Hausman 2010).

A causal generalisation can also be true without being instantiated: 'Eating one kilogram of uranium 235 causes death' (Hitchcock 1995: 236).[8] Moreover, claims about singular causation do not typically constitute evidence for generic causal hypotheses. To the contrary, most methods for establishing single-case causal claims require knowledge of the corresponding population-level claim (for history, see Scriven 1966; for law, Cranor 2011).

Clark Glymour's *bootstrapping account* (Glymour 1980) aims to improve upon some of the weaknesses in Hempel's theory. One of the problems of Hempel's satisfaction theory was that it doesn't let us use statements about observables provide evidence for hypotheses concerning unobservables. Glymour's account repairs this defect by allowing the use of *theory* in interpreting evidence. Roughly,

(**BOOT**) *e* is (supporting) evidence for *h* with respect to theory *t* if

(a) *e* and *t* entail an instance of *h*;

(b) there exists alternative evidence *e*´ such that *e*´ and *t* entail $\neg h$ in an inference analogous to that of (a).

Condition (b) is there to ensure that *e* plays any role in the derivation of *h*. (a) on its own can lead to a trivialisation of the condition when the evidence *e* plays no role in entailing *h* because to contains *h*.

(**BOOT**) doesn't easily apply to the support of *causal* hypotheses. Causal hypotheses are not simple generalisations from causal claims about individuals, as we have seen above. That *I* causes *D* in individual *i* may or may not mean that *I* causes *D* in the population *p* from which *i* was drawn. If *i* suffers from a rare genetic condition that makes wine poisonous for him, it may both be true that 'Drinking red wine in moderate amounts causes ill health in *i*' and that 'Drinking red wine in moderate amounts causes good health in *p*'.

Moreover, background 'theory' is never strong enough to entail, together with the evidence, an instance of the generalisation. Here 'theory' would refer to all the back-

[8] According to Nancy Cartwright for '*I* causes *D*' to be true it is enough that some *I*'s cause *D*'s (*e.g.*, Cartwright 1989). This has the somewhat awkward consequence that (for instance) both 'Seat belts save lives' and 'Seat belts kill' are true. Even if we accept this, this theory does not help with uninstantiated generic causal claims.

ground knowledge necessary to ensure that the evidence (such as 'Peter's cough stopped after taking the medicine') entails the instance of the hypotheses ('Peter's cough was relieved by the medicine'). This would mean we have made sure that there exists no other reason for which Peter's cough may disappear except the medicine. Of course, there are always open-ended lists of factors that may compete with a cause in bringing about an effect.[9]

I should mention that there is an approach in econometrics that seems to be represented well with Glymour's bootstrapping method: the structural or Cowles Commission approach. Here specific forms of econometric models are derived from theory and confirmed using correlations between variables of interest. Evidence plays a crucial role in deriving an instance of the hypothesis because it gives values to parameters. Confirmation is not trivialised because it is always possible that a parameter the hypothesis describes as positive or negative turns out to be zero.

Structural econometrics is very controversial within economics, however. The 'theory' that is used to interpret data is highly disputed, and so econometricians tend not to regard others' estimations (the estimations of those who use different bits of theory) as credible.[10] Thus, even if the bootstrap theory rationalises one approach within econometrics, little is won. Bootstrapping is generally inapplicable to the support of causal hypotheses.

Let me look at an alternative account, one implicitly given by Mill's methods. Mill's methods are of course not designed as a theory of evidence. However, they do address our two questions: What kinds of facts do we have to collect in order to support a (in Mill's as in our case, causal) hypothesis? What constitutes a good reason to believe or act on a hypothesis? Consider for instance the method of difference, which he describes as follows (Mill 1874[1843]):

> If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

To find out whether *I* causes *D*, Mill tells us to look for a situation in which *D* obtains and a second situation in which it does not obtain but which is otherwise identical in all respects except, possibly, one. If there is a factor with respect to which the two situations differ, then it is an effect or a cause or an indispensable part of the latter. The evidence is thus for a weaker disjunctive hypothesis but further evidence, for instance, about temporal order or an application of Mill's method of agreement, can rule out the unwanted alternatives.

Mill's theory belongs to the family of satisfaction theories because an instance of a (causal) generalisation confirms the generalisation. If, say, we find in a controlled experiment that, in the context, some factor *I* makes a difference to an outcome *D*, we

[9] Even if we were to weaken the relationship between evidence, theory and hypothesis instance to one of 'partial entailment', the problem would remain that causal hypotheses are not generalisations of instances of causal claims.

[10] For a discussion of the debate between structural econometrics and the opposing 'design-based' econometrics, see Reiss (forthcoming b).

not only infer that *I* caused *D in this case* but regard the causal generalisation *I*'s cause *D*'s as confirmed.[11]

In addition to the problem that generic causal hypotheses are not straightforward generalisations of singular causal claims, a main issue troubling Mill's theory is that it works only under ideal conditions. *If* we can find or create two situations that differ only in cause and effect, *then* we have a good reason to accept a causal hypothesis. But there aren't many hypotheses for which it is likely to find evidence of this kind. To begin with, no two situations are ever exactly alike (except with respect to cause and effect). Any application of the method requires judgements of causal relevance. Whether or not two situations in fact are 'relevantly' alike and how reliable our judgements concerning this matter are depends on the domain at hand and the state of our knowledge about it. Mill knew well that the method isn't universally applicable. Here is what he says about some of the domains in which we are interested here (Mill 1948 [1830]):

> But this can seldom be done in the moral sciences, owing to the immense multitude of the influencing circumstances, and our very scanty means of varying the experiment. Even in operating upon an individual mind... we cannot often obtain a crucial experiment. The effect, for example, of a particular circumstance in education, upon the formation of character, may be tried in a variety of cases, but we can hardly ever be certain that any two of those cases differ in all their circumstances except the solitary one of which we wish to estimate the influence. In how much greater a degree must this difficulty exist in the affairs of states, where even the number of recorded experiments is so scanty in comparison with the variety and multitude of the circumstances concerned in each. How, for example, can we obtain a crucial experiment on the effect of a restrictive commercial policy upon national wealth? We must find two nations alike in every other respect, or at least possessed, in a degree exactly equal, of everything which conduces to national opulence, and adopting exactly the same policy in all their other affairs, but differing in this only, that one of them adopts a system of commercial restrictions, and the other adopts free trade. This would be a decisive experiment, similar to those which we can almost always obtain in experimental physics. Doubtless this would be the most conclusive evidence of all if we could get it. But let any one consider how infinitely numerous and various are the circumstances which either directly or indirectly do or may influence the amount of the national wealth, and then ask himself what are the probabilities that in the longest revolution of ages two nations will be found, which agree, and can be shown to agree, in all those circumstances except one?

We need a theory of evidence that works for non-ideal cases, cases of the kind Mill describes here. The problem with Mill's methods is that they are completely silent about what we learn in non-ideal scenarios. When they apply, Mill's methods give us warrant of the highest degree, or as Mill calls it, 'the most conclusive evidence'. But surely we learn *something* from comparing two countries that are very much (but not completely) alike or from comparing many countries that differ with respect to a large range of (but not all) factors.

[11] Another way to classify Mill's theory would be as a forerunner of Mayo and Spanos' error-statistical approach (see below). After all, Mill's controlled experiments are designed to make sure that the same outcome would, in all likelihood, not have been produced if the causal hypothesis were false. That could only happen if another cause of *D* was present in the test situation and absent in the control, but overlooked. Controlled experiment are designed in order to avoid just that. What matters here are the characteristics of Mill's theory, not how we best classify it.

My remarks on the descriptive adequacy of Mill's theory are exactly the same as those concerning the hypothetico-deductivism. A strict reading of Mill's theory can hardly be descriptively adequate because it is not practicable. A looser reading of the theory can be regarded as underwriting controlled experiments and is therefore adequate to all those areas of science in which controlled experiments are used.

3.4 Error Statistics

Deborah Mayo and Aris Spanos's error statistical account of evidence is informative about what facts to gather in support of a hypothesis. The core concept in the account is that of a *severe test*. Mayo defines it as follows (Mayo 2000, S198; notation slightly altered for consistency with mine):

> Hypothesis *h* passes a severe test with *e* if (i) *e* fits *h* [for a suitable notion of fit or distance] and (ii) the test procedure *T* has a very low probability of producing a result that fits *h* as well as (or better than) *e* does, if *h* were false or incorrect.

The word 'fit' from the first clause is a statistical notion but Mayo intends something weaker. In other publications she and Spanos say 'agrees' instead of 'fits' (*e.g.*, Mayo and Spanos 2010: 22). Presumably, a correlation between *I* and *D*—recorded in an observational study or as a result of a randomised controlled study—fits or agrees with the hypothesis *h*: *I* causes *D*. The second clause is the crucial one. The test procedure must be designed in such a way that if *h* were false—*I* did not cause *D*—then the probability that it would produce a result like *e* would be very low. The error-statistical view of evidence thus distinguishes sharply between the methods by which evidence *e* is generated. An observational study T_O might not control for all common causes that may be responsible for the association between *I* and *D*. T_O therefore does not have a low probability that it produces a result that fits *h* as well as *e* (the correlation) does if a common cause is responsible for the association. By contrast, a well designed experiment T_E will control for this possibility and thus have a low probability to produce a result like *e*.

Only data produced by a virtually flawless procedure receives the honorific 'good evidence' according to Mayo (*op. cit.*):

> Data *e* produced by procedure *T* provides good evidence for hypothesis *h* to the extent that test *T* severely passes *h* with *e*.

Like Mill's, the error-statistical view collapses support and warrant into one concept—or at least Mayo doesn't tell us what to do with data that are not produced by severe tests. The problems are therefore analogous. If we are in the lucky position to know that *e* was produced by a severe test, the error-statistical view correctly identifies *e* as relevant to *h* and, equally correctly, as a good reason to infer *h*. However, a correlation recorded in an observational study, a non-severe test, would, incorrectly, not be regarded as evidence at all. And that is certainly mistaken, both normatively and descriptively.[12]

[12] Admittedly, Mayo defines only a sufficient condition for a severe test in the quotation above which suggests that there are other ways for a test to be severe than the one she describes. (Though at times she does include 'and only if', see for instance Mayo 2005). Moreover, in joint work with Aris Spanos,

3.5 Achinstein's Theory

In his *Book of Evidence* Peter Achinstein explicitly aims to address challenge of developing an account of evidence that is relevant to scientific practice. He identifies two main problems with the philosophical theories he examines—Bayesianism, hypothetico-deductivism and satisfaction theories à la Hempel 1945: they characterise a concept of evidence that is too weak for scientific purposes and they regard the evidential relationship as *a priori*.

We have seen in what senses these theories build on a weak concept of evidence: what these theories regard as evidence is at best support for a hypothesis, but it does not constitute a good reason to infer the hypothesis. To develop his own theory, Achinstein defines four concepts of evidence: potential, veridical, ES (epistemic-situation), and subjective. Potential evidence is the most basic and important concept. He characterises it as follows (Achinstein 2001, 170; notation slightly changed for consistency with mine):

(PE) *e* is potential evidence that *h*, given [background information] *b*, only if
1. Prob(there is an explanatory connection between *h* and *e*/*e*&*b*) > ½
2. *e* and *b* are true
3. *e* does not entail *h*.

There is an explanatory connection between *h* and *e* whenever either (i) *h* correctly explains why *e* is true, (ii) *e* correctly explains why *h* is true, or (iii) there is a hypothesis that correctly explains why both *h* and *e* are true. Veridical evidence adds the clause that (iv) *h* is true. ES evidence is relative to an epistemic situation and states that *e* is ES evidence whenever *e* is true and everyone in the epistemic situation believes that it is veridical evidence, and subjective evidence requires a subject *X* to believe that *e* is veridical evidence and take *e* as the reason to believe *h*.

Achinstein's theory is at best one of warrant. This is intended as Achinstein thinks of evidence as a 'good reason to believe'. He does not think that one needs a separate concept of relevant or supporting evidence (*ibid.*, 74):

> Accordingly, I reject the ambiguity response. Even if probabilists were correct in supposing that there is a sense of evidence that involves the idea of increase-in-strength-of-evidence, and even if the latter is connected to probability, it does not follow, and indeed is false, that any increase in probability is an increase in the strength of the evidence.

His main argument is that evidence is a threshold concept. For a hypothesis to have any acceptability or firmness, its probability must exceed some threshold. Just as adding one person to a group of people does not turn a non-crowd into a crowd, in-

she emphasises that 'test' may refer to a series of studies rather than a single, highly controlled procedure (Mayo and Spanos 2010, 25):

> Although it is convenient to continue to speak of a severe test T in the realm of substantive scientific inference... it should be emphasized that reference to "test T" may actually, and usually does, combine individual tests and inferences together; likewise, the data may combine results of several tests. To avoid confusion, it may be necessary to distinguish whether we have in mind several tests or a given test – a single data set or all information relevant to a given problem.

To my knowledge, this suggestion is never worked out, however.

crementally increasing the probability of a hypothesis does not constitute a (good or otherwise) reason to believe the hypothesis.

I fully agree that it 'is false, that any increase in probability is an increase in the strength of the evidence'. But it does not follow that there is no concept of evidentiary relevance or support. Tons of facts can be put on the table as supporting evidence for a hypothesis before a rational person is obliged or licensed or may feel encouraged to believe or act on the hypothesis. Supporting evidence is given by all the facts that are relevant to the assessment of a hypothesis. The outcome of this assessment is entirely independent of the status of the facts as relevant. And *this* concept of evidence is not a threshold concept. There is an intuitive sense in which adding relevant facts that speak in favour of a hypothesis strengthen it. However, the strength of evidence does not generally increase linearly with the number of facts that speak in favour of the hypothesis (perhaps weighed by the strength with which they speak in favour of it). Rather, there is a certain structure to evidential support, a structure which is missing from both Achinstein's and the Bayesian accounts.

Be that as it may, we should ask about the virtues of Achinstein's account as a theory of warrant. Achinstein adds the requirement that there be an explanatory connection between evidence and hypothesis in order to rule out cases where *e* is irrelevant to the assessment of the probability of *h* and the posterior of *h* is already high. The probability that Michael Jordan will not get pregnant (*h*) is high, given our background assumption that he is a male (*b*); whether or not he eats Wheaties (*e*) is irrelevant. And yet, Prob(*h* | *e*&*b*) > ½, which would mean that *e* is evidence for *h* under a pure high probability account (145ff.). Obviously, that Jordan eats Wheaties does not explain why won't get pregnant or vice versa, and so *e* is not evidence under Achinstein's account.

His theory therefore requires not that the posterior of *h* on *e* be greater than one half but the posterior of there being an explanatory connection between *h* and *e*, given *e* be greater than one half. This addition, however, introduces new difficulties.

It is often stated that causal relations explain correlations. There is certainly an informal sense in which this is true: *one reason* for two variables being correlated is that they are causally connected; the variables can be correlated *because* one causes the other. However, beyond this intuition it is hard to say precisely in which sense a causal relation explains a correlation. Correlations are mathematical constructs and therefore it is not the case that causal relations *cause* correlations. It is not clear, to say the least, that the hypothesis that *I* causes *D* is unifying. What are the diverse sets of explananda that can be derived from the hypothesis? If the causal relation is very robust in the sense that *I* causes *D* in many different populations and under very diverse sets of conditions, then the hypothesis may well be unifying, but it is entirely consistent with a causal hypothesis that the conditions under which it operates are highly restrictive and local. Further, there are many more reasons for which the two variables can be correlated: selection bias, certain statistical properties of the variables, conceptional, logical and mathematical relations between the variables, measurement error and so on. These different reasons are certainly not all equally unifying.

There is, to be sure, something like an inferential relationship between a causal hypothesis and a statement of a correlation: if we accept the causal hypothesis we can infer that the variables will (probably/likely/possibly) be correlated. Inferential relations are only sometimes explanatory, however. Observing a drop in the barometer reading I can infer that there (probably/likely/possibly) will be a storm but the drop in the barometer reading does not explain the storm. To find a criterion between explanatory and non-explanatory inferences was a problem logical empiricism struggled to solve for much of its existence, and I will not try to solve it here.

So there is at best an intuitive sense in which a causal hypothesis explains a correlation—the existence of a causal relation constitutes one reason among many for the existence of a correlation. Even if we accept this, the requirement that there be an explanatory connection between *h* and *e* is too strong. Evidence relevant to the assessment of a hypothesis is often very remote. Suppose you read a study purporting to show that *I* causes *D*. For all you know, the study was well designed, so if there really is a correlation between *I* and *D*—as the study reports—you will have a good reason that *I* does indeed cause *D*. However, it is part of your background knowledge that the team of researchers responsible for the study tends to be a bit sloppy when it comes to using spreadsheets, so you're not so certain that the correlation they report really exists in the population from which the data are drawn. But then you hear from an acquaintance that the team has taken criticisms of earlier episodes of sloppy spreadsheet calculations very seriously and installed a rigorous replication system where their results are independently calculated by two teams of graduate students. Now, in the present circumstances this piece of information will provide a good reason to infer the causal hypothesis and thus constitute warrant. But it neither explains nor is explained by the causal relation between *I* and *D*, nor is there another hypothesis that explains both this piece of information and the causal relation. The requirement of explanatory connectedness between evidence and hypothesis is too strong.[13]

Achinstein might think that his second clause helps with this problem. The clause requires that the evidential statement *e* and background assumptions *b* be true. After all, if *I* and *D* only appear to be correlated because of a coding error in the spreadsheet used, the statement '*I* and *D* are correlated' is not true and therefore not (potential) evidence according to the theory.

However, to require that the evidential statement *e* (and background assumptions *b*) be true introduces at least two new problems. First, neither evidential statements

[13] One might argue that the piece of information is not evidence as such but rather part of the background information *b* that helps to determine P('there is an explanatory connection between *h* and *e*' | *b*) > ½, where *h* is the causal hypothesis and *e* a sentence describing the correlation. (Thanks to an anonymous referee for this suggestion.) The problem with this reading of his theory is that Achinstein doesn't give us a criterion to determine what counts as relevant background knowledge. Certainly the explanatory connection does not give us relevance. Perhaps anything that helps to raise P('there is an explanatory connection between *h* and *e*' | *b*) to over ½? But how would we know this probability if we don't already have a concept of evidence (which Achinstein doesn't give)? At any rate, the piece of information described above is evidence but Achinstein's theory doesn't regard it as such.

nor background assumptions wear the label 'I am true' on their sleeves. For a theory of evidence to be practicable, we should only use conditions in the definition of evidence that are relatively readily ascertainable. If we don't, we might never be in the position to know that we do have evidence. While there is always the possibility of error in applying the definition, knowing that the conditions apply should not in principle be inaccessible. To ascertain that *e* and *b* are true is, however, beyond the average scientist's (or anyone else's) grasp.

Second, it is not always clear whether an evidential statement is true or not, even ignoring epistemic considerations. Once more, take correlations as an example. There is, to my knowledge, no generally accepted definition of what a correlation is. There are definitions of various correlation coefficients but it would be a mistake to think that two variables are correlated if and only if, say, the Pearson correlation coefficient is (significantly? at what level?) different from zero. Why use the Pearson coefficient and not any of the others? More substantially, there is a controversy over whether or not two non-stationary time series (*i.e.*, time series whose moments such as mean and variance change over time) are correlated. Kevin Hoover argues they are not, I argue that they are (Hoover 2003; Reiss 2007). The facts about which both parties agree are: (1) X_t and Y_t are two non-stationary time series, (2) the Pearson correlation coefficient $\varrho_{X,Y} \neq 0$; (3) X and Y are not causally connected. Is our evidential statement e = 'X, Y are correlated' true or false?

Something similar happens when the data are created by inadvertently conditioning on a common effect of the two variables through, for instance, selection. Suppose that I and D are causally and probabilistically independent in the general population. Suppose also that they have a common effect. Both I and D are conditions that make people see a doctor. If we look at data coming exclusively from doctors' records, we'll find I and D correlated. Thus, the two variables are uncontroversially correlated, though in the wrong population. One might try to solve this particular problem by requiring that the evidential statement contains information about the population ('I, D are correlated in the relevant population') but it will hardly be possible to operationalise the concept of 'relevant population'.

From the scientist's point of view it is insignificant whether evidential statements are true or not. If a study claims that two variables are correlated, then that's the evidence. There may be tens of reasons for the result other than a direct causal relation: confounding, selection bias, non-stationarity, mismeasurement, coding errors, fraud, to name but a few. To be justified in inferring the hypothesis presupposes having controlled for these errors. There may be a meaningful distinction between those errors that obtain when the correlation is genuine but not indicative of a direct causal relation (*e.g.*, confounding) and those that obtain when the correlation is not genuine to begin with (*e.g.*, mismeasurement, coding errors) but, as I have argued, it is somewhat blurry and not consequential for the inference from evidence to hypothesis.

To the extent that scientists use evidence that is not explanatorily connected with the hypothesis, Achinstein's account is not descriptively adequate. In favour of the account speaks its ability to deal with a range of non-ideal scenarios.

3.6 Cartwright's Argument Theory

Nancy Cartwright defends an 'Argument Theory' of evidence according to which (Cartwright forthcoming, 13-14):

> An empirical claim *e* is evidence for an empirical hypothesis *h* just in case *e* is an essential premise in a sound argument for *h*, that is, a valid argument with true premises.

This view on evidence is no doubt influenced by Cartwright's recent work on evidence-based medicine, policy and practice. Why, one might ask for instance, is a correlation produced by an randomised controlled trial (RCT) evidence for a causal hypothesis? Here is a sketch of an argument (Cartwright 2007, 13-14):

> To test 'T causes O' in φ via an RCT, we suppose that we study a test population φ all of whose members are governed by the same causal structure, CS, for O and which is described by a probability distribution P. P is defined over the event space $\{O, T, K_1, K_2, \ldots, K_n\}$, where each K_i is a state description over 'all other' causes of O except T. [Footnote omitted] The K_i are thus maximally causally homogeneous subpopulations of φ. Roughly
>
> - 'K_i is a state description over other causes' = K_i holds fixed all causes of O other than T.
> - 'Causal structure' = the network of causal pathways by which O can be produced, with their related strengths of efficacy.
>
> Then assume
>
> 1. *Probabilistic theory of causality.* T causes O in φ if $P(O/T\&K_i) > P(O/\neg T\&K_i)$ for some subpopulation K_i with $P(K_i) > 0$.
> 2. *Idealization.* In an ideal RCT for 'T causes O in φ', the K_i are distributed identically between the treatment and control groups.
>
> From 1 and 2 it follows that ideal RCTs are clinchers. If P(O) in treatment group > P(O) in the control group in an ideal RCT, then trivially by probability theory $P(O/T\&K_i) > P(O/\neg T\&K_i)$ for some K_i. Therefore: if P(O) in treatment group > P(O) in control group, T causes O in φ relative to CS, P.

According to the Argument Theory, a correlation between *T* and *O* is evidence for the hypothesis that *T* causes *O* only relative to an argument such as the above and the truth of its premisses. Formulated this way, the theory looks very harsh indeed. When would we *ever* be in the position to have evidence for a hypothesis? When would we be able to tell?

The paper from which the long quotation is taken distinguishes between 'clinchers' and 'vouchers' among methods for warranting causal claims. The former prove a hypothesis, given the assumptions (and are, consequently, narrow in their range of application), the latter speak in favour of the hypothesis without demonstrating it (and are broader in their range of application). An RCT is an example of a clincher, and so are certain econometric methods, Galilean experiments and derivation from established theory. Examples for vouchers are the hypothetico-deductive method (in its positivist, not its Popperian reading), qualitative comparative analysis, or 'looking for quantity and variety of evidence' (6).

The Argument Theory effectively denies that there are vouchers. Or, to put it more agreeably, the Argument Theory allows results produced by vouchers to count as evidence only to the extent that they have been converted into clinchers. How does one convert a voucher into a clincher? By adding strong inductive principles such as premiss 1 from the quote above. Suppose that we have recorded a correlation in an observational study. We might make the correlation vouch for a causal hypothesis hypothetico-deductively: the HD-method says that an observation we would expect to make were the hypothesis true speaks in favour of the hypothesis; if the causal hypothesis were true, we'd expect the variables to be correlated; thus, to record the correlation speaks in favour of the hypothesis.

Not according to the Argument Theory. Under that theory, we would have to write down an argument such as the following:

1.´ For any two variable *X* and *Y*, if *X* and *Y* are correlated, then they are 'causally connected'.

2.´ Two variables *X*, *Y* are causally connected if and only if *X* causes *Y*, *Y* causes *X* or a set of third factors *Z* causes both *X* and *Y*.

3. *I* and *D* are correlated.

4. *D* does not cause *I*.

5. There is no set of factors *Z* such that *Z* causes both *I* and *D*.

6. Therefore, *I* causes *D*.

The problem is that general principles such as 1. or 1.´ are false. A counterexample to 1.´ we've seen above: two non-stationary time series can be correlated and yet not causally connected. Perhaps we can avoid this problem by assuming weaker principles. After all, we only need a bridge between probability and causality *in this case*, not for any set of variables in any circumstance. Thus:

1.´´ If *I* and *D* are correlated, then they are 'causally connected',

where by *I* and *D* I am referring to the independent and dependent variable *in this case*. There are no reasons to believe that local principles such as 1.´´ could never be true. But what could be positive reasons to believe that they are true? Obviously, any reason to believe that any of the various non-causal accounts of correlations is absent. If we know that *I* and *D* are not correlated because they are non-stationary time series, because they are conceptually or logically or mathematically related or because they have been measured by conditioning on a common effect (and so on), then we have reason to believe that they are causally connected. According to the Argument Theory, to have evidence for a premiss such as 1.´´ means to have another argument with that statement as a conclusion. Perhaps:

7. If *I* and *D* are correlated, then they are 'causally connected' or there is a non-causal reason for the correlation.

8. No non-causal reason for a correlation applies to *I* and *D*.

9. Therefore, if *I* and *D* are correlated, then they are 'causally connected'.

The problem with premisses such as 8. (as well as 5.) is that the list of non-causal reasons for which two variables can be correlated is open-ended. Assuming these kinds of premisses for the purposes of arguing in favour of a hypothesis therefore always involves an inductive risk. No reformulation of episodes of inductive reasoning as formally deductively valid arguments could change that fact.

Cartwright demands that inductive arguments be reformulated as deductive ones in order to make explicit the principles on which these arguments are based—to force researchers think hard about reasons for using this or that principle in the given case and to know where potentially invalid steps in the inference lie. However, while the Argument Theory forces one to explicate one's inductive principles (such as 1., 1.´, 1.´´, 7. or 'If observed swans 1 through 333 have been white, then all swans are white'), it also hides important issues behind principles and statements about individuals assumed to be true. Take statement 5.: 'There is no set of factors *Z* such that *Z* causes both *I* and *D*.' Such a statement cannot be proved, just as the statement 'There are no unicorns' cannot proved. The best we can do is to rule out the set of known and relevant factors. What is known and what is relevant differs from case to case. The Argument Theory glosses over such case-specific differences.

How does the Argument Theory fare with respect to our four desiderata? First, support. Cartwright writes (Cartwright forthcoming, 14):

> To figure out whether *e* is evidence for *h*, the Argument Theory guides you to look for good arguments connecting *e* and *h*. Of course it doesn't tell you how to tell if an argument is good. But that's not in its job description. Coming up with an argument is part of the ordinary normal science job of scientific discovery. To check that it is valid, perhaps one needs a good logician or a good mathematician. To tell if the premises true, we employ the normal methods available in the paradigm in which we work for assessing the kinds of claims the premises make.

This is just like the Bayesian would answer: 'To figure out whether *e* is evidence for *h*, the Bayesian Theory guides you to look for statements *e* that are such that the probability of *h* on *e* is higher than the probability of *h*. Of course it doesn't tell you how to tell if a statement does raise the probability of *h*. But that's not in its job description. Coming up with statements that do is part of the ordinary normal science job of scientific discovery.' My arguments about Bayesianism therefore apply here too.

On the other hand, once we do have a good argument in favour of a hypothesis, we have in fact proved it (relative to the truth of the premisses) and therefore have a (somewhat coarse) way of assessing the hypothesis. Further, as long as we can come up with inductive principles that connect the evidence with the hypothesis, there are no a priori constraints about what can count as evidence. The theory therefore applies to non-ideal scenarios or any scenario at all. Scientists certainly use induction when reasoning about evidence without being able to formulate explicit inductive principles that can be used to write down deductive arguments every time. The Argument Theory is therefore at best an idealisation of practice or, more likely, to be understood as a revisionary rather than descriptive account.

4. *Analysis*

All the theories of evidence considered here play down at best but more often ignore the role the context of an inquiry plays in the determination of what kinds of facts are relevant to the assessment of a hypothesis and to how it is to be assessed. Let me discuss three contextual factors here: factual background commitments, the purpose of the inquiry, and normative commitments.

4.1 Factual Background Commitments

This is certainly a truism but in my view underappreciated by standard philosophical theories of evidence: what's evidence in favour (or against) a hypothesis is dependent on how the world works and our knowledge thereof. M's fingerprints on the murder weapon would not be evidence that M committed the murder if our fingers (as well as our hands, toes and feet) didn't have friction ridges or if the friction ridges weren't so varied as to allow individual identification. Nor would they be evidence if dactyloscopy hadn't been developed in the late 19th century. The same is true of evidence for scientific hypotheses. Relations of actual causation (such as killings) issue in typical markers and so do generic causal relations. That, for instance, causal relations are typically stable under intervention and can therefore be investigated experimentally is a contingent fact about causal relations that has to be discovered. Similarly, that a two-slit experiment produces the phenomenon of diffraction is evidence for the wave theory of light only because of facts about how other kinds of waves behave and our knowledge of these facts.

According to John Norton's 'material theory of induction', '*All inductions ultimately derive their licenses from facts pertinent to the matter of the induction*' (Norton 2003, 650; original emphasis). What Norton says here about induction is, subject to a caveat, true of evidence as well. Whether or not a thing is evidence for a hypothesis depends on facts pertinent to the matter of the hypothesis. The caveat is that whether or not a thing is evidence for a hypothesis does not *only* depend on those facts. It also depends on our knowledge of these facts and other contextual factors (which I will discuss below).

To be sure, some standard frameworks include places for background knowledge. Bayesianism, for instance, is more accurately formulated like this:

(**BAY′**) e is evidence for h given b if and only if $\text{Prob}(h \mid e, b) > \text{Prob}(h \mid b)$,

and hypothetico-deductivism like this:

(**HD′**) e is evidence for h given b if and only if h & b deductively entails e,

where b signifies background knowledge. However, in these and other theories it is never worked out just how background knowledge affects the status of a thing as evidence and, in particular, how changes in background assumptions can result in changes of the status of a thing as evidence.

Background knowledge in my sense will help with defining what the support of a hypothesis is, for instance, because it is informative about its empirical content.

4.2 Purpose

Another fairly trivial observation about the context of an inquiry is that the purpose of the inquiry makes an important difference to both relevance and hypothesis assessment. It is one thing to establish that someone committed the crime; quite another, to establish it in such a way as to convince a jury and comply with legal standards. An illegally taped telephone conversation may be compelling evidence in the former case but not constitute evidence at all in the latter; conversely, the defence might come up with certain outré alternative accounts of the defendant's behaviour for whose elimination evidence must be collected, facts that are likely to be entirely irrelevant outside the court.

How unusual alternative hypotheses can be for them to remain worthy of consideration is itself a question that the purpose of an inquiry helps to settle. If the aim is to establish a scientific hypothesis, there is usually no need to rule out sceptical alternatives such as evil-demon hypotheses. Michael Williams calls the injunction to ignore sceptical hypotheses in a scientific (or everyday) inquiry a 'methodological' constraint of the context because violating it would mean to change the nature of the debate (Williams 2001, 160). If, in the course of establishing whether *I* causes *D* we consider the hypotheses that the correlation between *I* and *D* was brought about by an evil demon, we no longer pursue a scientific question but a philosophical one. Similarly, a presupposition of any historical investigation is to suppress worries whether the Earth even existed five minutes ago. To allow that possibility would be to stop investigating historically (*ibid.*).

The audience to be addressed matters too. It is one thing, for instance, to establish the law of the free fall for oneself as it were, as Galileo did. It is quite another to establish it in a way that would be acceptable to the Aristotelian. By and large, Aristotelians did not accept Galileo's thought experiments, for instance, and would counter them with reports of observations of actual falls of bodies of different weights (Shea 1972, 11).[14]

Another sources of context-dependency that relates to the purpose of the investigation has to do with the ambiguity of scientific hypotheses. Take again causal hypotheses. A generic causal hypothesis of the form '*I* causes *D* in population *p*' is ambiguous in various ways. The following a some uncontroversial ways in which a causal hypothesis is ambiguous:

- *Component vs net effect.* A variable *I* may raise (in the causal sense) *D*'s probability along one route and lower it (also in the causal sense) along another. *I* is a component cause and preventer at the same time. *I*'s net effect on *D* may then be positive, negative or zero. Knowing *I*'s component and net effects can both be useful information, but in different contexts. Suppose *I* is smoking and *D* heart disease. Smoking is said to cause heart disease. Suppose also that smokers tend to exercise more, perhaps because they know of the negative (component!) effects of smoking and wish to compensate. Finally, suppose that the net effect is negative, because exercise is a strong preventer of heart disease. In this scenario, we can use smoking as a predictor of heart disease and, for instance, predict that as numbers of smokers decline, heart disease incidents should go up. But it would hardly be a prudent strategy to recommend to people that if they want to reduce their chances to get heart disease they should start to smoke. In this case the component effect is the

[14] Perhaps the Galilean case is one of evidentiary standards that vary with metaphysical background assumptions, in this case the assumption that reality is articulated in universal and stable modes about which we learn best when they are as free as possible of 'accidents' (local, idiosyncratic factors) and which was rejected by Aristotelians (McAllister 2004). A metaphysical background belief is a contextual feature too, of course.

appropriate quantity and the right recommendation would be to smoke less *and* exercise more often.[15]

- *Unanimous vs average effect.* In a randomised trial, we learn about the average effectiveness of treatments. When a treatment is effective on average across all subpopulations, it is possible that the effect has the opposite sign for some subpopulations. This cannot happen when the effect is 'unanimous', which means that it has the same sign for all subpopulations. If *I* has an average positive effect on *D*, then raising *I* will lead to an increase in *D* in the population. But this does not mean at all that giving *I* to an individual will have the beneficial effect, if that individual happens to be in one of the subpopulations where *I* has the opposite effect. Knowing *I*'s average effect on *D* is thus certainly useful for prediction, but it is controversial whether the average effect alone can underwrite population policies, and an average effect should certainly not be the only basis for an individual recommendation.
- *Necessary vs sufficient cause.* Causes almost always need other conditions to bring about their effects: matches do not light when struck unless oxygen is present in the air; venoms are toxic only in animals with certain genetic make-ups and when no antidote is present; development aid is effective only to the extent that the aid-receiving government is not too corrupt and the right kinds of socio-economic institutions are in place. The absence of each of these necessary causes (and equally, the absence of each of the other necessary conditions) will *explain* the absence of the effect. Only the whole amalgam of conditions will bring about the effect, however. If *bringing about an effect* is the purpose, we therefore need a sufficient or, as Mill says, a 'real cause' (Mill 1874[1843], Book 3, Chapter 5, §3).

It is not hard to see that the different interpretations of 'cause' in the causal hypothesis that are adequate in the light of the different purposes discussed above require different kinds of evidence (for details, see Reiss 2009, 2012).

4.3 Consequences and Normative Commitments

Scientific inquiry doesn't come for free. There are direct costs, ethical costs and opportunity costs. Even if it were the case that experimental studies are always more reliable than observational studies, this would not mean that the former are always preferable to the latter. Experimental studies are nearly always more financially and ethically costly than analogous observational studies, and if the consequences of making a mistake in the assessment of a hypothesis are small, it may well be better to forgo the additional reliability.

That value judgements affect scientific research is widely argued (for an overview of the arguments, see Reiss and Sprenger forthcoming). The gathering of evidence (which includes decisions about what information to seek, how to seek and how long to seek) and the assessment of hypothesis in the light of the gathered evidence are

[15] To the extent that that is possible. If the exercise variable cannot be manipulated independently, smoking more may well be the best strategy to prevent heart disease.

examples of kinds of scientific activity that are affected by value judgements. Importantly, these kinds of scientific activity are affected by value judgements in a way that the epistemic dimensions cannot neatly be separated from the pragmatic dimensions of a problem (*cf.* Kitcher 2011, 31ff.). Clearly, then, it is norms and values that help to determine the degree of warrant a hypothesis enjoys.

Nancy Cartwright sees the entanglement of the epistemic and pragmatic dimensions of hypothesis assessment clearly. It is worth quoting her in full (Cartwright forthcoming, 13):

> Consider: You are about to endorse a claim to a graduate student whom you know is readily influenced by you and is considering taking a position in a research group that uses this claim as a central pillar for its research. Before endorsing this claim in these circumstances, you should consider the evidence for it. You should also consider the abilities of the research team that propose to follow it up, the opinion of your colleagues about the evidence and what it shows, the talents of the student, the chances that she will end up with publishable papers even if the research program does not produce its promised results, and so forth. These issues will not separate nicely, as we might have hoped, to afford a two-stage deliberation: first wear your scientist's hat to estimate the degree to which you are justified in 'accepting' the claim; then consider how justified you are in using a claim with that degree of warrant in the way proposed. Rather you must consider the issues all together in one fell swoop. And you should consider them. What you say to the student matters to her life, so you should take pains to ensure that what you do is justified. But that is not an exclusively scientific enterprise.

Her 'Argument Theory' of evidence appears not to be suited to the joint consideration of epistemic and pragmatic issues 'in one fell swoop' she describes in this passage. Even if the theory is only meant to be one of relevance and not of assessment or 'warrant', one cannot help but feel that a hypothesis which is supported by a deductively valid argument with true premisses has as much warrant as there can be.

5. Conclusions

My last remark concerning Cartwright's Argument Theory generalises to all theories of evidence considered here. Most or all of what I said in Section 4 are well-known observations about the importance of context in scientific inquiry. But while well-known, standard theories of evidence (and non-standard theories such as Cartwright's) fail to acknowledge them and integrate them into the theoretical framework. What we need is a contextualist theory of evidence.

REFERENCES

Achinstein, P. 1995. Are Empirical Evidence Claims A Priori? *British Journal for Philosophy of Science* 46(4): 447-473.

—. 2001. *The Book of Evidence*. Oxford: Oxford University Press.

Carnap, R. 1947. On the Application of Inductive Logic. *Philosophy and Phenomenological Research* 8(1): 133-148.

Cartwright, N. 2001. What's Wrong With Bayes' Nets? *Monist* 84(2): 242-264.

—. 2007. Are RCTs the Gold Standard? *BioSocieties* 2(2): 11-20.

Cranor, C. 2011. *Legally Poisoned: How the Law Puts us at Risk from Toxicants*. Cambridge, MA: Harvard University Press.

Glymour, C. 1980. *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Hausman, D. 2010. Probabilistic Causality and Causal Generalizations. In *The Place of Probability in Science*, eds. E. Eells and J. Fetzer. Dordrecht: Springer.

Hempel, C. 1945. Studies in the Logic of Confirmation (I.). *Mind* 54(213): 1-26.

Hitchcock, C. 1995. Mishap at Reichenbach Fall: Singular vs. General Causation. *Philosophical Studies* 78(3): 257-291.

Hoover, K. 2003. Nonstationary Time-Series, Cointegration, and the Principle of the Common Cause. *British Journal for the Philosophy of Science* 54: 527-551.

Howson, C. and P. Urbach. 1993. *Scientific Reasoning: The Bayesian Approach*. Chicago, IL: Open Court.

Kitcher, P. 2011. *Science in a Democratic Society*. Amherst, NY: Prometheus Books.

Mayo, D. 1996. Error and the Growth of Experimental Knowledge. Chicago: University of Chicago Press.

—. 2000. Experimental Practice and an Error Statistical Account of Evidence. *Philosophy of Science* 67(Proceedings): S193-207.

—. 2005. Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses. In *Scientific Evidence: Philosophical Theories and Applications*, ed. P. Achinstein. Baltimore, MD: The Johns Hopkins University Press: 95-128.

— and A. Spanos. 2010. *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. Cambridge: Cambridge University Press.

McAllister, J. 2004. Thought Experiments and the Belief in Phenomena. *Philosophy of Science* 71: 1164-1175.

Mill, J. S. 1874 [1843]. *A System of Logic*. New York, NY: Harper.

—. 1948 [1830]. *Essays On Some Unsettled Questions of Political Economy*. London: Parker.

Norton, J. 2003. A Material Theory of Induction. *Philosophy of Science* 70(4): 647-670.

—. 2010. *A Survey of Inductive Generalization*. University of Pittsburgh.

—. 2011. Challenges to Bayesian Confirmation Theory. *Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics,* eds. Prasanta S. Bandyopadhyay and Malcolm R. Forster. Dordrecht: Elsevier.

Rescher, N. 1958. A Theory of Evidence. *Philosophy of Science* 25(1): 83-94.

Reiss, J. 2007. Time Series, Nonsense Correlations and the Principle of the Common Cause. In *Causality and Probability in the Sciences*, eds. F. Russo and J. Williamson. London: College Publications: 179-196.

—. 2009. Causation in the Social Sciences: Evidence, Inference, Purpose. *Philosophy of the Social Sciences* 39(1): 20-40.

—. 2012. Third Time's a Charm: Wittgensteinian Pluralisms and Causation. In *Causality in the Sciences*, eds. P. McKay Illari, F. Russo and J. Williamson. Oxford: Oxford University Press: 907-927.

—. Forthcoming a. A Theory of Inferential Judgement. Manuscript. Durham, Durham University.

—. Forthcoming b. Two Approaches to Reasoning From Evidence or What Econometrics Can Learn from Biomedical Research. Manuscript. Durham, Durham University.

— and J. Sprenger. Forthcoming. Scientific Objectivity. In *Stanford Encyclopedia of Philosophy*, ed. E. Zalta. Stanford, CA.

Salmon, W. 1975. Confirmation and Relevance. In *Induction, Probability, and Confirmation*, eds. G. Maxwell and R. Anderson. Minneapolis, MN: University of Minnesota Press: 3-36.

Scriven, M. 1966. Causes, Connections and Conditions in History. In *Philosophical Analysis and History*, ed. W. Dray. New York, NY: Harper and Row: 238-264.

Shea, W. 1972. *Galileo's Intellectual Revolution*. London: Macmillan.

Williams, M. 2001. *Problems of Knowledge*. Oxford: Oxford University Press.

Williamson, J. 2010. *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.

JULIAN REISS is Professor of Philosophy at Durham University. He has a degree in economics and finance from the University of St Gallen and a PhD in philosophy from the London School of Economics. His main research interests are methodologies of the sciences (especially causality and causal inference, models, simulations and thought experiments, and counterfactuals), philosophy of economics, and science and values. He is the author of *Error in Economics: Towards a More Evidence-Based Methodology* (2008), *Philosophy of Economics: A Contemporary Introduction* (2013), and over 40 papers in leading philosophy and social science journals and edited collections.

ADDRESS: Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, UK. E-mail: julian.reiss@durham.ac.uk