A Pragmatist Theory of Evidence

Julian Reiss*†

Two approaches to evidential reasoning compete in the biomedical and social sciences: the experimental and the pragmatist. Whereas experimentalism has received considerable philosophical analysis and support since the times of Bacon and Mill (and continues to enjoy attention and support in very recent work on causation and evidence), pragmatism about evidence has been neither articulated nor defended. The overall aim is to fill this gap and develop a theory that articulates the latter. The main ideas of the theory will be illustrated and supported by a case study on the smoking/lung cancer controversy in the 1950s.

1. Introduction. There are two paradigms of reasoning from evidence at work in the biomedical and social sciences (cf. Parascandola 2004). There is, on the one hand, the experimental paradigm, according to which randomized experiments constitute the 'gold standard' of evidence and all other methods are assessed in terms of how closely they resemble the gold standard. The experimental paradigm is currently dominant in all the domains labeled 'evidence-based', which include parts of medicine, dentistry, nursing, psychology, education, social policy, and criminal justice, but also parts of development economics.

There is, on the other hand, the pragmatist paradigm, according to which scientific claims are inferred, using pragmatic criteria, from diverse bodies of evidence that may but need not include experiments. Many scientists across the biomedical and social sciences subscribe to the pragmatist paradigm, albeit usually less candidly than the proponents of experimentalism.

Received December 2013; revised November 2014.

*To contact the author, please write to: Department of Philosophy, Durham University, Durham DH1 3HN, UK; e-mail: julian.reiss@durham.ac.uk.

†A previous draft of this paper was discussed with the Centre for Humanities Engaging Science and Society (CHESS) research group at Durham University and improved considerably. Thanks also to Bert Leuridan for comments. Financial support from projects FFI2008-01580/Consolider Ingenio CSD2009-0056 and FFI2011-23267 of the Spanish Ministry of Science and Innovation is gratefully acknowledged.

Philosophy of Science, 82 (July 2015) pp. 341–362. 0031-8248/2015/8203-0001$10.00

The experimental paradigm has received considerable philosophical analysis and support since the times of Bacon and Mill. Indeed, Mill's methods are best understood as accounts of controlled experimentation, and more recent work on evidence and causality can be used to underwrite randomized controlled trials (Mayo 1996; Woodward 2003; Cartwright 2007). Even the philosophical literature that takes a critical stance toward evidence-based medicine, policy, and practice tends to focus on the virtues and vices of randomized experimentation.

The pragmatist paradigm is much harder to articulate and defend. Among other things, the paradigm seems to raise more questions than it answers: What are the supposedly 'pragmatic criteria'? What is a diverse 'body' of evidence? Just how 'diverse' does it have to be? And how do we know what is to be included (as evidence) if there's no standard against which to judge? The aim of this article is to answer these questions. More broadly speaking, I aim to develop a theory of evidence that articulates the pragmatist paradigm and serves as an alternative to the experimentalist paradigm that currently dominates the discussion.

As the pragmatist theory of evidence is to serve as an alternative to a paradigm that takes randomized experimentation as the gold standard, I will focus on scientific domains where randomized experiments can be and are frequently employed. This includes the domains mentioned above but excludes all those domains where controlled experiments are effectively epistemic engines, such as large parts of physics and chemistry and basic/in vitro research in the biomedical sciences. I shall also exclude historical sciences such as cosmology, astronomy, astrophysics, geology, palaeontology, and archaeology. I do believe that the proposed account can be extended, but I will leave the extension to future work.

2. Preliminaries. Before developing the theory, I need to prepare the ground by distinguishing between two concepts of evidence, both of which are needed in a satisfactory theory of evidence, laying out a number of desiderata a good theory should satisfy, and describing a number of caveats for this article.

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 mean that *e* is a 'mark' or 'sign' or 'symptom' of the hypothesis being true, that *e* is a piece of evidence for *h*. A correlation, say, between two variables *I* and *D* is evidence in this sense for the hypothesis *h*: '*I* causes *D*.'[1] To learn that *I* and *D* are correlated supports (speaks in favor of) the hypothesis without yet constituting a reason to infer the hypothesis, even a

1. I use 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.

weak one. 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 'proof' or 'warrant' that *h* or that *e* constitutes a '(weak, strong, etc.) reason to infer' *h*, or that *e* is 'body of evidence' for *h*. It is harder to find an unequivocal and simple example for this type, but suppose that the correlation between *I* and *D* was established in a well-designed randomized trial, treatment and control group are known to be balanced, greatest care was taken to avoid coding and measurement error, and so on; then this body of knowledge together constitutes what I will call 'warranting evidence' or 'warrant'.

The distinction I have in mind can be illustrated by a kind of interrogation that is familiar from murder mysteries on TV. When the detective investigating the murder case asks someone who is in one way or another related to the murder (by being involved with the victim, being at the crime scene, or what have you) for an alibi, the putative suspect frequently gets defensive and replies, "Do you believe I have anything to do with the murder? I would never . . . !" Detectives then often counter, "I do not believe anything," and then, "I only collect facts," or "I have to exclude this possibility," or "I have to ask this." If a putative suspect does not have an alibi, this is a piece of information that speaks in favor of (or does not speak against and is relevant to) the hypothesis that the putative suspect was the murderer. As such it has nothing to do whatsoever with belief. It can, in a different process (which often occurs at a later stage but can also be simultaneous), lead to a belief revision and inference to a hypothesis. However, to collect facts and to make up one's mind (i.e., to infer a hypothesis) are two different activities. 'Support' relates to the collection of facts; 'warrant,' to making up one's mind. 'Evidence', unfortunately, conflates the two.

A good theory of evidence should explicate both support and warrant. We need, on the one hand, criteria or guidelines that tell us what kinds of facts we have to collect in order to evaluate a hypothesis; we need to know what facts are relevant to the hypothesis. We need, on the other hand, criteria or guidelines that tell us how to assess the hypothesis, given the facts we've collected in its support, or, conversely, criteria or guidelines that tell us how much support of what kind we need in order to achieve a given degree of warrant. We require criteria or guidelines that translate between knowledge of the facts relevant to a hypothesis and judgments about the hypothesis.

A theory of evidence that didn't tell us about relevance would be impracticable; a theory that didn't tell us about assessment would not be useful. Here, then, is a first desideratum for us: the theory should be a theory of both support and warrant.

Further, it is clear that warrant comes in degrees. We can have better and less good reasons to infer a hypothesis; a hypothesis can be more or less warranted. Thus, the second desideratum is that the theory is informative about the degree to which evidence warrants a hypothesis. There is no presumption here that degrees of warrant are probabilities, only that the theory allows hypotheses to be weakly ordered, at least sometimes, with respect to warrant.

Lastly, for a theory to be useful it should tell us about warrant in both ideal and nonideal epistemic circumstances. Consider the following ideal theory of evidence:

> (**ITE**) Hypothesis *h* is strongly warranted if and only if the results *e* of a flawless randomized controlled trial fit *h* (on a suitable notion of 'fit').

This would presumably get the judgment right in cases where it does apply, but it would seldom if ever apply. A good theory continues to provide useful information when randomization fails or cannot be done, when hypotheses are established by means of observational studies, when knowledge of the phenomena of interest is limited, and so on. In sum, a good theory of evidence

1. distinguishes support and warrant;
2. provides an account of evidential support;
3. provides an account of warrant that allows warrant to come in degrees; and
4. applies to nonideal circumstances typical of science in practice.

Like John Norton, I maintain that justification for inductive inferences is local and material (e.g., Norton 2003). One cannot say very much about evidence and how it supports hypotheses at a level of high generality. Here I will focus on a specific type of scientific hypotheses in a relatively small range of domains. My examples concern scientific hypotheses expressing

- causal relations;
- between type-level variables (rather than token-level or relations of actual causation); and
- in those parts of the biomedical and social sciences where randomized experiments can be and are frequently used.

3. Support: The Eliminativist Hypothetico-Contextualist (EHC) Framework. The pragmatist theory of evidence proposed here is remotely related to the hypothetico-deductive theory of confirmation. Hypothetico-deductivism (HD), once defended by prominent philosophers of science (Ayer 1936/1971; Popper 1963; Hempel 1966), has had bad press in philosophy for over 30 years. Clark Glymour once called it "hopeless" (Glymour 1980).

I will argue in what follows that the philosophers' obituaries have been premature and that we should not give up on the basic idea behind the theory. My own framework therefore retains the 'hypothetico' of HD. Philosophical critics are mistaken because they focus on the logical properties of the 'deductive' part of the theory. I will argue that the fault lies with this interpretation of the theory, not with the core idea behind the theory itself. While my account differs considerably from standard HD as described by philosophers, it captures the kind of reasoning we find in scientific practice (in the areas on which I focus here) well.

HD holds that that which is deductively entailed by a hypothesis provides support for it. More precisely,

> (**Standard-HD**) A statement *e* provides support for hypothesis *h* if and only if *h* (possibly in conjunction with suitable background knowledge) deductively entails *e*.

To be deductively entailed by a hypothesis is, however, neither necessary nor sufficient for providing support. Hypotheses typically do not entail anything about the specific data sets that are used in their support. For example, statistical hypotheses do not entail any statements describing particular data sets; causal hypotheses do not entail statements about correlations or invariance. The first example is straightforward. Suppose we observe a series of 50 heads in 50 coin tosses. This is certainly support for the hypothesis that the coin is biased. But the hypothesis "This coin is biased" does not entail a description of this particular series of outcomes or any other.

Similarly, there is no guarantee that causal relations induce correlations in the relevant data sets. If, to rehearse a standard counterexample, *I* causes *D* via two different routes, say, directly and via an intermediary *R* as in figure 1, *I* can be marginally uncorrelated with *D* even if the variables are correlated conditional on *R*.[2] *R* might be a variable we don't know about or one that's not measurable.

2. I should mention that this example is ruled out by what is called the "faithfulness condition" (Spirtes, Glymour, and Scheines 2000) or "stability condition" (Pearl 2000). Spirtes et al. argue that an exact cancellation of influences through the two different routes has Lebesgue measure zero. Their argument has been criticized, however, for two principal reasons. First, exact cancellations are often what we try to achieve with policies (Hoover 2001). To the extent that our policies are successful, we should expect cancellations to occur. Second, real-world methods never allow us to determine whether or not an exact cancellation has occurred anyway. Our philosophy should be relevant to science as it is practiced, and from an empirical point of view there is no way of telling whether an exact or rather a near-exact canceling has occurred (Cartwright 1999). Faithfulness and stability are extremely powerful assumptions where they work, but we should not bet on their being universally true axioms.

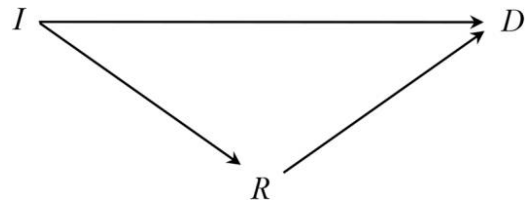

Figure 1. *I* causes *D* through two different routes.

At any rate, in our example the support is a marginal correlation between *I* and *D*. And a statement describing the marginal correlation is definitely not entailed by the causal hypothesis.

Conversely, any statement entails itself, but no self-respecting biomedical or social scientist would take the truth of a hypothesis as support for itself.

'Set of deductive consequences' is, however, only one way to understand the empirical content of a hypothesis. I propose to regard the relationship between hypothesis and its support as inductive rather than deductive. In particular, to determine the support of a hypothesis, we have to ask what patterns in the data we would expect to hold if the hypothesis were true, given our understanding of how the world works. To use a mundane example, murderers often leave traces on murder weapons. But not to find a suspect's fingerprints on the murder weapon does not demonstrate the falsehood of the hypothesis because fingerprints on the murder weapon may fail to be detected for any number of reasons: the murderer wore gloves, she wiped them off, she threw the weapon into a river, it rained, a cat licked them off, our fingerprint detection technology failed, a member of our forensic team received a bribe and lied about the result, and so on. Nevertheless, background knowledge concerning how murders happen entitles one to expect a suspect's traces on a murder weapon under the supposition that he or she is the murderer and therefore, if found, constitutes support.[3]

3. Let me make two remarks at this point. First, there is no sharp distinction between background knowledge, on the one hand, and evidence or support, on the other. In the context of a given case we might distinguish between entrenched beliefs and new information that was produced in order to assess the hypothesis at hand, but there is no presumption to the effect that background knowledge must be true or cannot be challenged. To the contrary, every factual claim that is used in the assessment of a hypothesis can in principle be contested. I discuss the issue of the circumstances under and the extent to which these claims should be challenged in some detail below. Second, the notion of support is not unlike the Bayesian notion of 'partial entailment', albeit without the probabilities. I cannot make a case against Bayesianism here in any detail. Let me just say that there are no physical probabilities for conditional statements such as "*X* leaves fingerprints on the murder weapon given *X* is the murderer" or "*I* and *D* are

Similarly, a correlation is the kind of thing one is entitled to expect to find if a causal hypothesis were true even though the absence of a correlation does not prove the falsity of the hypothesis. Let us then characterize support as follows:

> (**S**) *e* provides support for a hypothesis *h* if and only if *e* is a pattern in the data we are entitled to expect to obtain under the supposition that *h* is true (see Hempel 1966, 6).

A second problem of standard-HD has been referred to as the "problem of alternative hypotheses" (see Mayo 1996, chap. 6): if *e* supports *h*, it may also support alternatives that are incompatible with *h*; *e* on its own does not discriminate between *h* and the alternatives the supposition of whose truth also entitles us to expect *e* to obtain. A correlation between two variables *I* and *D* may support the causal hypothesis *h* "*I* causes *D*," but it may also support h' ("*D* causes *I*"), h'' ("A common factor *C* causes both *I* and *D*"), h''' ("The correlation between causally independent variables *I* and *D* was induced by conditioning on a common effect *E*"), and many others.

A straightforward solution to this problem is to postulate that support for a hypothesis is of two kinds: direct support, e_d, which pertains to the hypothesis of interest; and indirect support, e_i, which pertains to the elimination of alternative hypotheses. So far we have only looked at direct support. Indirect support provides another element in 'EHC': eliminativism.

Indirect support is given by patterns in the data that are incompatible with the truth of an alternative hypothesis. A suspect's fingerprints on the murder weapon are direct support for the hypothesis that the suspect killed the victim. A second suspect's alibi is indirect support because it helps to eliminate the hypothesis that the second suspect did it. Likewise, if a correlation provides direct support for a causal hypothesis, a study that shows that no common cause could be responsible for the correlation provides indirect support for the hypothesis. Let us then define:

> (**S-d**) e_d provides direct support for a hypothesis *h* if and only if e_d is a pattern in the data we are entitled to expect to obtain under the supposition that *h* is true.

> (**S-i**) e_i provides indirect support for a hypothesis *h* if and only if e_i is a pattern in the data that is incompatible with what we are entitled to expect to obtain under the supposition of the truth of one of *h*'s alternative hypotheses h', h'', h''', and so on.

---

correlated given that *I* causes *D*," to assume sharp subjective probabilities is hopelessly unrealistic and misleading, and to assume vague probabilities is to give up most of the advantages of Bayesianism. See Norton (2011) for an elaborate discussion. Possibility and plausibility are the modalities adequate for evidential reasoning, not probability.

Definition S-i is in fact somewhat ambiguous. In what sense are h', h'', h''', and so on, alternatives to h? In the present context they are alternative accounts of the evidence in favor of h. A suspect's accidental arrival at the crime scene and his picking up the murder weapon from the cold body is an alternative account for finding his fingerprints on the weapon (the latter of which speaks in favor of the initial hypothesis). Selection bias is an alternative account for a correlation (the latter of which speaks in favor of the initial hypothesis).

The ambiguity is that there are alternative accounts for all pieces of support, not just for the direct support. A piece of indirect support is a second suspect's alibi. But an alibi is never ascertained with full certainty. Rather, that the second suspect has an alibi is itself inferred from what she says, what others have observed, CCTV recordings, gas station receipts, credit card records, and so on. We thus have different pieces of indirect support that help to ascertain that the second suspect does in fact have an alibi, which rules her out as a perpetrator and thereby supports the initial hypothesis. But, of course, each of these pieces of indirect support also comes with alternative accounts. If she in fact did it, that's a good reason for saying that she was in a restaurant with her girlfriend at the time of the crime. If the girlfriend confirms this, their friendship or other kind of involvement may account for her testimony. The gas station receipt could be someone else's and the credit card record from the restaurant faked. Additional indirect evidence serves to rule out these possibilities.

Thus, each piece of indirect support can itself be accounted for by alternative hypotheses. If, say, a common-cause hypothesis is an alternative account of the correlation and a study that shows that there is no common cause that could account for the correlation is the indirect support, then there are alternative accounts for the results of this study. These too must be eliminated by further indirect support. This leads to the following, amended definition of indirect support:

> (**S-i***) e_i provides indirect support for a hypothesis h if and only if e_i is a pattern in the data that is incompatible with what we are entitled to expect to obtain under the supposition that (a) an alternative hypothesis able to account for h's direct support is true or (b) an alternative hypothesis able to account for h's prior indirect support is true.

Definition S-i* is not circular despite the occurrence of "indirect support" on the left and on the right of the "if and only if." The lowest-level indirect support is defined in terms of direct support. Higher-level indirect support is defined in terms of lower-level indirect support. However, there can be an infinite regress, namely, when all higher-level pieces of indirect support continue to have alternative hypotheses.

We have now 'solved' two problems of standard-HD by introducing four new problems: (1) How do we know what we are entitled to expect to ob-

tain under the supposition of the truth of a hypothesis? (2) How do we know what are the alternatives to *h* and which alternatives of a potentially infinite set of possible alternatives to consider? (3) How are alternative hypotheses eliminated? (4) How is the infinite regress in definition S-i* stopped? As we will see, the third element of the EHC framework, the context of a causal inquiry, will provide the answers.

4. Causal Inquiries in Context. The notion of expectation employed here is a contextual one. It is the context of a causal inquiry, itself given by background knowledge about how the world works, the nature and purpose of the inquiry, and certain normative commitments, that answers these questions. Let us then turn to an analysis of the contributions of context for each of the four questions raised above.

4.1. The Empirical Content of Causal Hypotheses. In previous work I have argued that a problem with standard theories of causation—probabilistic, regularity, interventionist, and process or mechanism—is that they mistake evidence for whether or not a causal relation is present for the relation itself or for constituting the meaning of causal claims. These are verificationist theories of causation and therefore suffer from the standard objections to verificationism (Reiss 2012a). In the present context, however, the verificationism of these theories is just what we need to determine the empirical content of a hypothesis. When we ask what we'd expect to find if the causal hypothesis "*I* causes *D*" were true, the standard theories of causation provide the following answers:[4]

- a correlation between *I* and *D*;
- *D*'s changing after an intervention on *I*;
- *I*'s being a necessary or sufficient condition or both for *D*, or *I*'s being an insufficient but nonredundant part of an unnecessary but sufficient (INUS) condition for *D*;
- a continuous process from *I* to *D*;
- a mechanism for the causal relation between *I* and *D*.

These expectations stem simply from general background knowledge about how the world works. We know that causal relations typically issue in cor-

4. I omitted the counterfactual theory here, which is a sixth 'standard' theory of causation. While I do believe that counterfactual claims about the value of *D* that would have obtained if the value of *I* had been different can constitute support for causal hypotheses, their relation to causal claims is rather involved, and they are themselves highly theoretical and require causal background knowledge to be established (Reiss 2009b, 2012b). To tease these complex relationships apart would require more space than I have here and distract from the overall line of the argument.

relations and regularities and help to bring about change through interventions. In the biomedical and social sciences we also know that causes typically do not produce their outcomes across spatiotemporal gaps and are often 'structured' in the sense of being dependent on underlying systems that are made up of varieties of mechanisms.

The last two items require some further comments. Correlations, changes in one variable following an intervention in another and one variable's being a necessary, sufficient, or INUS condition for another, are patterns in the data that, while strictly unobservable, are fairly readily verifiable, given the data (though see the remarks about coding errors below). This is not the case for claims about processes and mechanisms. Claims about processes or mechanisms are themselves established hypothetico-contextually. When we hypothesize that *I* causes *D* and that *I* causes *D* through a process or that a mechanism is responsible for the relation between *I* and *D*, we can make further hypotheses about the mode of action of the process or mechanism. Each hypothesis about one of the parts of the process or mechanism will license certain expectations about patterns in the data that should obtain were the hypothesis true, and finding these patterns (and other patterns that are incompatible with alternative hypotheses about the process or mechanism) will establish the hypothesis. Conjoining a number of such hypotheses will form a complex hypothesis about a process or mechanism, which in turn constitutes support for the original causal hypothesis.

Generally speaking, we have learned over time how causal relations behave, both at a high level of abstraction and concerning more specific causal relations in specific contexts. Whereas just 100 years ago causality was tightly wedded to determinism, we have more recently become accustomed to probabilistic causality. Thus, whereas a century ago we'd have expected that an effect must happen if its cause had (and we could use this expectation to rule out a factor as a cause if its effect does not happen despite its occurrence), for the most part we now expect causal relation to issue, at best, in correlations. Similarly, control over phenomena is one of the main purposes of learning causal relations since at least Bacon. However, the Lucas critique (Lucas 1976) has taught us that at least in economics we cannot always rely on causal relations for policy (Reiss 2008). It is background knowledge like this that determines the empirical content of causal hypotheses.

Everything said so far pertains to causal inquiries very generally. Focusing on more narrow types of inquiry or specific inquiries provides further information about what and what not to expect. To keep the discussion brief, I will focus my remarks mainly on a single case study, the controversy surrounding the hypothesis that smoking causes lung cancer in the 1950s.

4.2. Considering Alternatives. The direct support of the smoking/lung cancer hypothesis consisted in correlations recorded mainly in retrospective

case-control studies, but by the mid-1950s also early results of a prospective study (Doll and Hill 1956). The main alternative account for a correlation is that the correlation is spurious. But 'spurious correlation' is ambiguous, and the way the term is often used is misleading.

Literally speaking, when a correlation is said to be spurious, one means that the correlation is not genuine but merely apparent. This can happen for a variety of reasons. One source is the inadvertent conditioning on the wrong variable. Suppose that *I* and *D* are independent, dichotomous variables and one's data collection consists only of individuals where either I = true or D = true. If so, then *I* and *D* will be correlated in the data set but not in the general population.[5] The same happens when one conditions on a joint effect. Other reasons for correlations being spurious include mismeasurement, coding errors, sloppiness in keeping records, deliberate fraud, and so on.

'Spurious correlation' is more frequently but misleadingly used to refer to a confounded causal relation. Here *I* and *D* are genuinely correlated, but the correlation is due to a common cause or causality running in the opposite direction from *D* to *I*.

A third case obtains when statistical properties of time series induce correlations that cannot be causally explained. This is, for instance, the case when two time series monotonically increase (Sober 2001) and, more generally speaking, when the two time series are nonstationary (Hoover 2003). Correlations induced by properties of time series cannot readily be classified as either 'spurious' or 'confounded'.[6]

In general, an empirical reason is required for taking an alternative account of the direct support (or prior indirect support) to be relevant. In the context of a scientific inquiry it would be inappropriate to advance a general skeptical alternative, such as an evil-demon hypothesis (see Goldman 1976, 775). Among the empirical reasons are generic reasons that pertain to all inquires of a given type and case-specific reasons. When correlations are

5. Define $B \equiv I$ v D. *I* and *D* are probabilistically independent: Prob($D \mid I$) = Prob(D). Examining a data set that consists only of individuals for which either *I* or *D* is true is equivalent to conditioning on *B*. Conditional on *B*, *I* and *D* are probabilistically dependent: Prob($D \mid I, B$) ≠ Prob($D \mid B$). This is called Berkson's paradox. That the data are selected in this way is not always conspicuous to the researcher.

6. This is therefore an interesting case for theories of evidence that require the evidential statement that *e* be true. Whether or not two nonstationary time series (i.e., time series whose moments such as mean and variance change over time) are correlated is controversial. Kevin Hoover argues that they are not; I argue that they are (Hoover 2003; Reiss 2007). The facts about which both parties agree are as follows: if X_t and Y_t are the two nonstationary time series, (1) the Pearson correlation coefficient $r_{X,Y} \neq 0$, and (2) *X* and *Y* are not causally connected. So if our evidential statement e = "*X*, *Y* are correlated," is *e* true or false?

recorded in observational studies, background knowledge tells us that selection bias is always a relevant alternative. Similarly, Berkson's paradox is a relevant alternative when the studies draw on hospitalized patients.

In the smoking/lung cancer case, both types of alternatives were relevant. One important alternative was Ronald Fisher's 'constitutional hypothesis', according to which a common genetic factor is responsible for the correlation. Joseph Berkson pointed out that there is a danger of bias if the control group is not selected in such a way as to represent (with respect to smoking habits) the general population, which includes the lung cancer patients—which was the case in the retrospective studies that were drawn on hospitalized patients. Mismeasurement (in this case, 'diagnostic error') too was an alternative that was known to possibly account for the observed correlation. If many of those who died of other diseases, such as tuberculosis, were classified as lung cancer cases, a spurious association could be generated. At the time it was known, for instance, that an error in tuberculosis diagnosis of only 11% could account for the entire recorded increase in lung cancer (Gilliam 1955). This was, then, certainly a relevant alternative.

A researcher can also show an alternative to be relevant by presenting direct support for it. Fisher supported his views about the smoking/lung cancer link with a study demonstrating that monozygotic twins are more likely to be alike with respect to their smoking behavior than dizygotic twins, even if they were separated at birth (Fisher 1958). This is just what we would expect if genetics played a role in determining smoking behavior. An alternative for which there is direct support I will call a 'salient' alternative.

That cancer susceptibility was partly based on genetics was well known at the time. The psychologist Hans Eysenck and his colleagues showed that smoking was related to extroversion, which in turn had a genetic component (Eysenck et al. 1960). A noteworthy feature of that study was that it showed a dose-response effect: the more extroverted a person, the more she smokes.

4.3. Eliminating Alternatives. Alternatives are eliminated by pointing to patterns in the data that are incompatible with what we would expect to be the case were an alternative true. The following are some patterns in the data researchers have used in order to eliminate alternative hypotheses in the smoking/lung cancer case:

- *Confounding*. A number of patterns in the data were appealed to to eliminate alternative causal accounts. For example, there is a large dose-response effect. Moderate smokers have a ninefold greater risk of developing lung cancer than nonsmokers, while over-two-pack-a-day smokers have at least a 60-fold greater risk. There was no known

genetic factor that could produce such a strong effect. A study of lung cancer and blood groups (which were known to have a genetic basis) showed a difference of only 27% (Fisher 1958). Further, there is a strong stopping effect in that individuals who discontinue smoking have a much lower risk of developing the disease. The genetic factor cannot therefore be constant over an individual's lifetime, which is highly implausible given what was known about genetics (Cornfield et al. 1959). Another piece of indirect support was that lung cancer prevalence in males increased long before it did in females. If a genetic factor were appealed to in order to explain this observation, there would have to have been a mutation in males first and a few decades later in females, a pattern that had not previously been observed (Cornfield et al. 1959).

- *Spurious correlation*. In 1951, Doll and Hill sent questionnaires to 40,000 British doctors asking about smoking behavior and recorded mortality subsequently. First results from this study became available in the mid-1950s. These confirmed a dramatic increase in lung cancer risk among smokers but could not be accounted for by Berkson's paradox (Doll and Hill 1956).
- *Diagnostic error*. In the mid-1950s there were good reasons to believe that numerous death cases were misclassified. However, the misclassification hypothesis cannot explain micropatterns in the data. Assuming that lung cancer prevalence was stable over time would mean a diagnostic error of only 3% among those 35–44 years of age but 59% among those 75 years or older. Similarly, there would be different rates of diagnostic error for men and women (Gilliam 1955). It is certainly possible that there are different error rates in different patient groups, but that the error in older patients should be an order of magnitude larger than that in younger patients is extremely unlikely.

Values do and should play a role in the decision whether or not to reject an alternative in light of incompatible patterns in the data. If little hinges on the decision, we may keep entertaining an alternative even in light of dramatic indirect support. If, by contrast, a decision is likely to have significant welfare consequences (as, of course, was the case with respect to alternatives to the causal hypothesis in the smoking/lung cancer case), the standards for rejecting an alternative should be lower. There are no strict rules, however, that map the cost of maintaining a false alternative to a threshold of 'strength of support' beyond which it becomes strictly irrational to do so.

It is therefore important to note that no amount of incompatible information can 'prove' an alternative wrong. It would not necessarily be irrational to continue to maintain that the constitutional hypothesis is correct

in light of a large dose-response effect—perhaps the smoking/lung cancer gene has a very peculiar mode of action. The rejection of an alternative always remains a judgment. Direct support and indirect support suggest a certain decision, but alternative decisions are possible and often defensible. Over half a century on, we may be inclined to think that those on the "right" side of the controversy had objectively better reasons than "those who were wrong." However, what one finds is "extremely well-written and cogent papers that might have become textbook classics for their impeccable logic and clear exposition of data and argument if only the authors had been on the right side" (Vandenbroucke 1989, 3). Support and logic by themselves do not compel a decision one way or another.

4.4. Ending the Regress. Each piece of indirect support has itself alternative hypotheses able to account for it. The study that shows that genetic factors can account for only 27% of cancer susceptibility may itself be subject to all sorts of biases, confounding, mismeasurement, error, and fraud. If we tried to rule out every one of these possibilities, we would never reach a stage where we could accept any hypothesis.

One suggestion that has been made is that epistemic trust helps to determine when to stop (Hardwig 1991). It would be impossible to control for all the potential alternatives; thus, if we didn't trust others, there would be no scientific knowledge. If a study claims that there is a certain pattern in the data, such as an association between two variables, as a general rule, we take this as a fact. We presume that if we were to replicate the study on the same data set, our investigations would yield the same result. We think that this is so because scientists take a reasonable amount of care when they make public assertions and because peer review constitutes a safeguard against errors.

It would be quite naive, however, to hope that epistemic trust can do all the work all the time. We don't have to appeal to scandals such as that about Vioxx in pharmaceutical research (Biddle 2007) or AusterityGate in economics (Reiss 2014) to see that. On the one hand, there are general statistical reasons to believe that "most published research findings are false" (Ioannidis 2005, e124). On the other hand, oftentimes there will be more specific reasons to mistrust particular findings or claims.

In such environments it is hard to take the bulk of published research results at face value. Nevertheless, we sometimes have to do that, or there would be no scientific progress. The following are some pragmatic guidelines that can help end the regress. The first is a general, philosophically motivated rule.

- *Default entitlement: as a default rule, scientists are entitled to each other's claims. They should probe claims only when there are domain-*

or case-specific reasons to do so. Justification in the sciences can be said to have what Robert Brandom calls a "default and challenge structure" (1994, 177). Scientists are entitled to each other's claims in the absence of appropriate reasons to think that they are not so entitled. When entitlements are challenged, the reasons given must be relevant in the context of a given causal inquiry. I understand claims broadly to include the main study results but also claims about raw data, as well as the protocols and methods used.

When there are relevant reasons to think that previous results should be probed, the following examples for supporting guidelines may help. In each case the guideline was at work in the smoking/lung cancer case, but it can independently be motivated and defended. The list is, of course, not meant to be exhaustive of the kinds of rules scientists use to eliminate alternative hypotheses.

- *Effect size*: *the larger the effect size a study reports, the smaller the need for probing the result*. Large effects can be a great help to the elimination of alternative explanations because alternatives become intolerably implausible. This criterion has limitations: it works only with some kinds of alternatives (e.g., not if fraud is suspected) and only in some circumstances (namely, when effect sizes are predicted and large), but it can help greatly where it works.
- *Manner and timing of the effect*: *the more specifically the manner and timing of the effect match the expectation, the smaller the need for probing the result*. Like effect size, the timing and manner of the effect can also be of great help with the elimination of alternative accounts. We may expect some pattern in the data on the assumption of a given alternative at some relatively abstract level of description, but, with luck, not at a more microscopic level. If smoking causes lung cancer, we expect more frequent smokers to have a higher risk, we expect stopping to have a beneficial effect, we expect the cancer to develop some time after an individual has taken up smoking rather than immediately, and so on.
- *Study characteristics: the smaller the number of background assumptions that are needed to derive a study result and the smaller the inferential gap between data and result, the smaller the need for probing the result*. There are large differences in the number and kind of inferences made between studies. Many involve highly sophisticated statistical techniques and shaky background assumptions. Others proceed on the grounds of well-entrenched procedures that have been around for decades or even centuries. Yet others may simply report summaries of the data in the form of histograms and tables, or even the

raw data themselves. All study results and all aspects of an individual study can in principle have alternative accounts. However, to the extent that claims involve a minimum amount or unproblematic inferences, and unless there are overwhelming reasons to believe otherwise, these claims can be (tentatively) accepted without further probing.

- *Economic and other normative considerations: take into account economic and other costs and benefits when deciding to stop or continue probing the indirect support for a hypothesis*. Causal inquiry does not come for free. There are direct, opportunity, and ethical costs. These costs have to be traded off against the benefits of reducing uncertainty. The benefits of reducing uncertainty consist in the reduced chance of accepting a false or rejecting a true hypothesis. There are no strict rules on how to optimize the trade-off, and people holding different values will differ in their assessments. What is clear, however, is that a reasonable trade-off will seldom entail an indefinite continuation of challenging the indirect support for a hypothesis.

These rules helped to resolve the smoking/lung cancer controversy fairly quickly. Researchers noted the parallel rise in cigarette consumption and lung cancer and began to investigate the relationship only in the 1930s. By the early 1950s, prospective studies were under way, and by the mid-1950s, a large part of the medical community was convinced of the carcinogenicity of cigarette smoke. Here are some of the facts that played a role in forming the consensus:

- *The effect is massive*. In 1956 Doll and Hill calculated that smokers of 25 or more cigarettes per day increased their odds of dying from lung cancer by a factor of about 24 (Doll and Hill 1956). By the end of the 1950s, data suggested that that factor could be as high as 60 (Cornfield et al. 1959).
- *The manner and timing of the effect are hard to account for by other hypotheses*. Lung cancer rates in the United States went up before they did in Canada, in parallel with the difference in smoking patterns between the two countries. This is hard to account for by a genetic modification. A genetic factor cannot account for the stopping effect. Other environmental factors cannot account for the sex differences in smoking behavior and cancer epidemiology. There is a large association between pipe and cigar smoking and cancer of the buccal cavity and larynx but not cancer of the lung.
- *Many studies used came as close to epistemic bedrock as it gets*. Gilliam (1955) effectively ruled out the ‘diagnostic error’ hypothesis by simply arranging mortality statistics according to age and sex. No mathematics or statistical technique was involved here other than drawing simple averages.

- *There is a widely shared norm that public health should address the fundamental causes of disease and aim to prevent adverse health outcomes*. Few researchers working in cancer epidemiology in the 1950s were motivated primarily by a commitment to smokers' enjoyment or the profits of the tobacco industry.[7] To form a consensus view concerning the dangers of cigarette smoke has costs in the form of reduced enjoyment (at least some smokers will give up in response), increased worry, the health consequences of increased worry, and the financial losses of all those involved in the production, marketing, and selling of cigarettes. These obtain quite independently of the truth of the hypothesis. If the hypothesis is true, but only if it is true, it has benefits in the form of a reduced health burden due to smoking. If there hadn't been a normative consensus on values—that the uncertain benefits outweigh the costs—it would have been a lot harder to form the epistemic consensus.

5. Warrant: Counting Eliminated Alternatives. The account of warrant I propose follows the EHC framework developed for support. Accordingly, a scientific hypothesis is warranted to the extent that (*a*) it has direct support and (*b*) alternative accounts of the direct support and indirect support have been eliminated. It is straightforward, then, to define different 'grades' of warrant. I propose to define four grades: proof, strong warrant, moderate warrant, and weak warrant. Table 1 shows how they are defined.

Calling warrant of the highest grade 'proof' is consistent with the scientific use of the term. For example, as early as 1953 Richard Doll wrote about the smoking/lung cancer link, "The results [described in this paper] amount, I believe to proof that smoking is a cause of bronchial carcinoma" (1953, 585). This concept of proof should, of course, not be confused with the mathematicians' and logicians' concept. In particular, to have proof for *h* does not entail that *h* must be true, given the support. It can always be the case that an alternative has been overlooked or that an alternative that has been eliminated should not have been. So the concept is one of empirical or inductive, not deductive, proof.[8]

The number of alternative accounts that have been eliminated is responsible for the strength of warrant. Salient alternatives, that is, alternatives for which there exists direct support, contribute more to the strength of the warrant than nonsalient alternatives because a true alternative is more likely to

7. Naomi Oreskes and Erik Conway's 'merchants of doubt' target beliefs about the adverse health consequences of secondhand smoke, not smoking itself (see Oreskes and Conway 2010).

8. An important issue concerns the (possible) existence of hitherto-unconceived alternatives (see Stanford 2006). Perhaps we shouldn't call a hypothesis proved if we have not

TABLE 1. Different Grades of Warrant

Grade	Name	Direct Support plus Indirect Support That . . .
1	Proof	Eliminates all (relevant) alternative accounts
2	Strong warrant	Eliminates all salient alternative accounts and some that are nonsalient
3	Moderate warrant	Eliminates most alternatives, including some that are salient
4	Weak warrant	Eliminates some alternative accounts

leave traces in the data than a false alternative. To have exactly three grades of warrant short of proof is, of course, arbitrary, but it is consistent with scientific practice, for instance, at the International Agency for Research on Cancer (see IARC 2006).

6. Experiments, Instruments, and Pragmatism. I began this article by distinguishing two approaches to reasoning from evidence in the biomedical and social sciences: the experimentalist and the pragmatist. Now that I have articulated the latter, what can we say about the relation between the two? The main difference, as I see it, is their mode of justification. Experimentalists are methodological foundationalists. They believe that some results are produced by methods that are intrinsically reliable and therefore epistemically basic. The epistemically basic method is a well-designed and well-executed randomized experiment.

Like other foundationalists, experimentalists have to address two fundamental issues (see Williams 2001, 85): First, how does one explain that the chosen kinds of methods are regarded as intrinsically reliable? Second, how can the success of other methods (those that are not intrinsically reliable) be explained with reference to the basic methods? One way to answer the first question is to underwrite the method with a theory of the nature of causality in such a way that the method can be shown to produce reliable results. Mill's view of causes as INUS conditions can be understood this way (as underwriting his methods of agreement and difference; see Mackie 1980, chap. 3 and appendix), and so can Woodward's theory of causation as invariance under intervention (as underwriting experiments, espe-

---

ruled out all relevant alternatives, whether already put forward or as yet unconceived. I'm willing to bite the bullet with respect to this issue. Proof is a contextual matter, and if no one has been able to come up with a plausible alternative account, it is not relevant in the given context. This may, of course, lead to cases where a hypothesis comes out as proved, and yet the true hypothesis hasn't even been conceived yet. As long as everyone understands that proof is a contextual and fallible matter, this doesn't seem to be too problematic. What has been proved today can be revised tomorrow on the basis of new findings, new technologies, new ideas. This happens all the time.

cially randomized experiments; see Woodward 2003). While I don't think that these are successful theories of causality (e.g., Reiss 2009a), let us, for the sake of the argument, suppose that these defenses work. What about the second issue?

The problem is that once experiments are regarded as epistemically basic because intrinsically reliable, it becomes very hard to explain why data produced by other methods should be able to support causal claims. If (ideal) experiments are the 'gold standard' of evidence, why are observational studies that record correlations among variables on which no intervention has taken place evidence at all? Why demographic trends? Why a study that records gender- and age-specific patterns in mortality data? To be sure, some methods resemble experiments. The definition of an instrumental variable in econometrics, for instance, is very similar to the definition of an 'intervention' in Woodward's theory (Reiss 2005). One can also show that randomization is an instrumental variable (Heckman 1996). However, most things resemble most other things in one respect or another, and experimentalism is silent about the kinds of respect in which a method must resemble an experiment in order to be able to support a causal claim. It is silent, too, about the extent to which dissimilarities should be punished ("At what level of dissimilarity is a method mere silver standard or bronze, and when is it rubbish?"). Evidence-based medicine and practice provide clumsy 'hierarchies of evidence' but no explanations of why a hierarchy should be thus and not otherwise.

The pragmatist theory of evidence proposed here has no trouble explaining the success of experiments and instrumental-variable studies. Both (well-designed) experiments and (well-designed) instrumental-variable studies are often reliable because they eliminate a host of alternatives—all alternative causal hypotheses—in one fell swoop. It can also explain failure when it occurs. Even a well-designed (natural or field) experiment can deliver botched results when variables are poorly measured, coding errors are made, computer programs are implemented sloppily, or data are inappropriately pooled. Experimentation as such cannot protect from these errors, and the theory proposed here highlights that all relevant alternatives (to the extent that cost-benefit considerations mandate it) should be eliminated.

One might counter at this point that the contrast I draw between experimentalism and pragmatism is exaggerated—that experimentalists are pragmatists at heart, except that they have a narrow(er) understanding of what good evidence is. I disagree, but I cannot give a full-blown empirical analysis of what scientists really believe here. Let me instead point out the following: (1) Even if this critic is right, the pragmatist alternative still needs to be articulated—and this is what this article has aimed to do. (2) The pragmatist theory proposed here can explain why randomized trials are successful where they are and describe the conditions under which they can be

expected to be successful. Simply declaring them to be the 'gold standard' does not provide such an explanation. (3) The remaining, important difference is that there is no gold standard of evidence whatsoever in the proposed account. Whoever regards certain kinds of experiment as the standard of evidence takes a starting point in the method used. The account proposed here instead begins with the hypothesis and inquires about what kinds of facts we need to collect or learn in order to be entitled to infer the hypothesis. That these facts can sometimes be learned in an experiment is trivially true but, according to this account, contingent and of no deeper significance for the justification of inferences.

7. Conclusions. Let me end by returning to the questions concerning the pragmatist paradigm with which I began. The EHC framework presented here gives the following answers:

> *Relevance*. How do we know what is to be included as evidence (in the sense of support) in the assessment of a hypothesis? There is no fully general answer. Evidence for a hypothesis is given by what we are entitled to expect under the supposition of the truth of the hypothesis. What we are entitled to expect is given by background knowledge about how the world works. Relevant to the assessment of a hypothesis is not only evidence in favor or against the hypothesis but also evidence in favor or against relevant alternative hypotheses.
> *Body of evidence*. The body of evidence is given by the totality of the (direct and indirect) support for a hypothesis. To what extent a hypothesis is warranted is determined on the basis of the totality of its support.
> *Diversity of evidence*. The body of evidence for a hypothesis is diverse in two senses. First, there is the distinction between direct and indirect support. A hypothesis cannot be warranted unless supported by both. Second, the indirect support, which is used to eliminate alternative hypotheses, will be as diverse as the alternative hypotheses it helps to eliminate. It is quite a different thing to show that the existence of a common cause is unlikely than it is to show that variables have been measured correctly or that peer review ensures that the chance of encountering coding or programming errors is low.
> *Pragmatic criteria*. Pragmatic criteria to address these issues have been discussed throughout section 4 (for instance, that knowledge about correlations; changes under intervention; necessary, sufficient, or INUS conditions; and processes/mechanisms constitutes direct support for causal hypotheses; what kinds of alternative hypotheses to consider; that researchers are entitled to other researchers' results unless there are good reasons to think that there is no such entitlements; and so on).

This article aims to contribute to a growing body of literature on evidence in the social and biomedical sciences. Unlike the earlier literature in the Carnapian and Bayesian traditions, the more recent work takes scientific practice a lot more seriously, both in terms of its greater use of knowledge about the conditions under which science is practiced and in terms of its goal to develop insights that are relevant to practicing scientists. The specific contribution I hope to make is to provide a realistic framework—a framework that applies to epistemic conditions that are nonideal—for thinking about evidence across the biomedical and social sciences within which more specific questions, such as about whether or not both mechanistic evidence and probabilistic evidence are required to establish a causal hypothesis (e.g., Russo and Williamson 2007), what the role of basic science is in evidence-based medicine (e.g., La Caze 2011), or how to interpret hierarchies of evidence (Borgerson 2009), can be addressed fruitfully. Whether the framework delivers on this promise is, alas, a matter for future research.

REFERENCES

Ayer, Alfred. 1936/1971. *Language, Truth and Logic*. Repr. London: Penguin.

Biddle, Justin. 2007. "Lessons from the Vioxx Debacle: What the Privatization of Science Can Teach Us about Social Epistemology." *Social Epistemology* 21 (1): 21–39.

Borgerson, Kirstin. 2009. "Valuing Evidence: Bias and the Evidence Hierarchy of Evidence-Based Medicine." *Perspectives in Biology and Medicine* 52 (2): 218–33.

Brandom, Robert. 1994. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Cartwright, Nancy. 1999. *The Dappled World*. Cambridge: Cambridge University Press.

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

Cornfield, Jerome, William Haenszel, Cuyler Hammond, Abraham Lilienfield, Michael Shimkin, and Ernst Wynder. 1959. "Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions." *Journal of the National Cancer Institute* 22:173–203.

Doll, Richard. 1953. "Bronchial Carcinoma: Incidence and Aetiology." *British Medical Journal* 2 (4836): 585–90.

Doll, Richard, and Austin Bradford Hill. 1956. "Lung Cancer and Other Causes of Death in Relation to Smoking: A Second Report on the Mortality of British Doctors." *British Medical Journal* 2 (5001): 1071–81.

Eysenck, Hans, Mollie Tarrant, Myra Woolf, and L. England. 1960. "Smoking and Personality." *British Medical Journal* 1 (5184): 1456–60.

Fisher, Ronald A. 1958. "Cancer and Smoking." *Nature* 182:596.

Gilliam, Alexander. 1955. "Trends of Mortality Attributed to Carcinoma of the Lung: Possible Effects of Faulty Certification of Deaths due to Other Respiratory Diseases." *Cancer* 8:1130–36.

Glymour, Clark. 1980. "Discussion: Hypothetico-Deductivism Is Hopeless." *Philosophy of Science* 47:322–25.

Goldman, Alvin. 1976. "Discrimination and Perceptual Knowledge." *Journal of Philosophy* 73 (20): 771–91.

Hardwig, John. 1991. "The Role of Trust in Knowledge." *Journal of Philosophy* 88 (12): 693–708.

Heckman, James. 1996. "Randomization as an Instrumental Variable." *Review of Economics and Statistics* 78 (2): 336–41.

Hempel, Carl. 1966. *The Philosophy of Natural Science*. Upper Saddle River, NJ: Prentice Hall.

Hoover, Kevin. 2001. *Causality in Macroeconomics*. Cambridge: Cambridge University Press.
———. 2003. "Nonstationary Time-Series, Cointegration, and the Principle of the Common Cause." *British Journal for the Philosophy of Science* 54:527–51.
IARC. 2006. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans: Preamble*. Lyon: International Agency for Research on Cancer.
Ioannidis, John. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124.
La Caze, Adam. 2011. "The Role of Basic Science in Evidence-Based Medicine." *Biology and Philosophy* 26 (1): 81–98.
Lucas, Robert. 1976. "Economic Policy Evaluation: A Critique." *Carnegie-Rochester Series on Public Policy* 1:19–46.
Mackie, John. 1980. *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.
Mayo, Deborah. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.
Norton, John. 2003. "A Material Theory of Induction." *Philosophy of Science* 70 (4): 647–70.
———. 2011. "Challenges to Bayesian Confirmation Theory." In *Philosophy of Statistics*, ed. P. Bandyopadhyay and M. Forster. Dordrecht: Elsevier.
Oreskes, Naomi, and Erik Conway. 2010. *Merchants of Doubt*. New York: Bloomsbury.
Parascandola, Mark. 2004. "Two Approaches to Etiology: The Debate over Smoking and Lung Cancer in the 1950s." *Endeavour* 28 (2): 81–86.
Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
Popper, Karl. 1963. *Conjectures and Refutations*. London: Routledge.
Reiss, Julian. 2005. "Causal Instrumental Variables and Interventions." *Philosophy of Science* 72 (Proceedings): 964–76.
———. 2007. "Time Series, Nonsense Correlations and the Principle of the Common Cause." In *Causality and Probability in the Sciences*, ed. F. Russo and J. Williamson, 179–96. London: College Publications.
———. 2008. *Error in Economics: Towards a More Evidence-Based Methodology.* London: Routledge.
———. 2009a. "Causation in the Social Sciences: Evidence, Inference, Purpose." *Philosophy of the Social Sciences* 39 (1): 20–40.
———. 2009b. "Counterfactuals, Thought Experiments and Singular Causal Analysis in History." *Philosophy of Science* 76:712–23.
———. 2012a. "Causation in the Sciences: An Inferentialist Account." *Studies in History and Philosophy of Biology and Biomedical Science* 43 (4): 769–77.
———. 2012b. "Counterfactuals." In *Oxford Handbook of the Philosophy of Social Science*, ed. H. Kincaid, 154–83. Oxford: Oxford University Press.
———. 2014. "Struggling over the Soul of Economics: Objectivity versus Expertise." In *Experts and Consensus in Social Science*, ed. C. Martini and M. Boumans. Cham: Springer.
Rescher, Nicholas. 1958. "A Theory of Evidence." *Philosophy of Science* 25 (1): 83–94.
Russo, Federica, and Jon Williamson. 2007. "Interpreting Causality in the Health Sciences." *International Studies in the Philosophy of Science* 21 (2): 157–70.
Salmon, Wesley. 1975. "Confirmation and Relevance." In *Induction, Probability, and Confirmation*, ed. G. Maxwell and R. Anderson, 3–36. Minneapolis: University of Minnesota Press.
Sober, Elliott. 2001. "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause." *British Journal for the Philosophy of Science* 52:331–46.
Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.
Stanford, P. Kyle. 2006. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.
Vandenbroucke, Jan P. 1989. "Those Who Were Wrong." *American Journal of Epidemiology* 130 (1): 3–5.
Williams, Michael. 2001. *Problems of Knowledge*. Oxford: Oxford University Press.
Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.