Inquiry and Evidence:
A Pragmatist Philosophy of Scientific Practice*

Matthew J. Brown†

December 18, 2018

Abstract

In the first part of this paper, I will sketch the main features of traditional models of evidence, indicating idealizations in such models that I regard as doing more harm than good. I will then proceed to elaborate on an alternative model of evidence that is functionalist, complex, dynamic, and contextual, which I will call *dynamic evidential functionalism*. I will demonstrate its application to an illuminating example of scientific inquiry, and defend it from some likely objections. In the second part, I will use that alternative to solve a variety of classic and contemporary problems in the literature on scientific evidence having to do with the empirical basis of science and the use of evidence in public policy.

Keywords: evidence, experimenter's regress, robustness, theory-ladenness, evidence-based policy, pragmatism

*My thanks to Nancy Carwright, Paul Churchland, Wayne Martin, Jacob Stegenga, the UCSD Philosophy of Science Reading Group, and participants of *Philosophy of Scientific Experimentation: A Challenge to Philosophy of Science* at the University of Pittsburgh Center for Philosophy of Science for comments on earlier versions of this project. §6 owes much to many able critics of the position I've laid out so far through personal and online conversation, notably Jacob Stegenga and several commenters on the philosophy of science blog *It's Only a Theory*, including Thomas Basbøll, Greg Frost-Arnold, Gabriele Contessa, and Eric Winsberg. The original discussions on that blog can be found there in several entries under the tag "evidence" from October 2009. An extract from Part I was published in 2015 in *Metaphilosophy* as "The Functional Complexity of Scientific Evidence."

†Center for Values in Medicine, Science, and Technology, School of Arts and Humanities, The University of Texas at Dallas, 800 W. Campbell Rd, JO 31, Richardson, TX 75080, mattbrown@utdallas.edu

Failure to institute a logic based inclusively and exclusively upon the operations of inquiry has enormous cultural consequences.

- John Dewey[1]

1 Introduction

Several problems in the contemporary philosophy of scientific evidence—the experimenter's regress,[2] concerns about discordant evidence,[3] worries about the importance of "robust" evidence from different types of sources,[4] attempts to better understand the nature and role of scientific experimentation vis a vis theory,[5] and questions about "evidence for use" as distinct from evidence for theories or hypotheses[6]—are dependent on a commonly assumed but radically impoverished model of evidence (or better, a family of such models). This model is impoverished in that it ignores the temporal dynamics of inquiry within which evidence plays a role, as well as the variety of functional roles for evidence within that dynamic process. Since the problems are the result of the features of the model (rather than problems of evidence as such), many attempts to solve the problems amount to evasions, to patches that generate even further problems, and so on. What is needed is a deep and systematic rethinking of the basic model of evidence underlying the various approaches to evidence that are current today.

Features of the traditional model are often held implicitly, causing frustration not only amongst philosophers, but also in some areas of the social and medical sciences as well as policy-making which have be influenced by that model. Scientists and policy-makers now find themselves in quandaries about how to rate evidence and how to combine it from multiple sources. Setting a framework of "evidence-based policy" is one of the latest practical problem about evidence to arise at the interface of science and policy.

In this paper, I will describe the main features of the traditional model(s) of evidence, indicating the idealizations of that model which I regard as doing more harm than good. I will then proceed to outline an alternative model of evidence – *dynamical evidential functionalism* (DEF); on the DEF-model, evidence is

(a) *Functionalist* – Evidence is defined by its functional role(s) within a scientific inquiry.

[1] Dewey (1938, *Logic*, (LW 12: 527))
[2] Collins (1992); Franklin (1994); Godin and Gingras (2002)
[3] Franklin (2002); Stegenga (2009)
[4] Culp (1994); Stegenga (2009); Hey (2015)
[5] Franklin (2005); Karaca (2013); Colaço (2018)
[6] Cartwright (2006)

(b) *Complex/Multi-functional* – Evidence plays a number of different functional roles, irreducible to any particular role.

(c) *Dynamical* – Scientific inquiries are processes with a beginning, middle, and end, and this dynamical structure is relevant for understanding the roles of evidence.

(d) *Contextual* – Evidence is relative to the context of the particular scientific inquiry in which it functions, and the aims and social context of the inquirers.

In order to lay out the model, I will first have to lay out the larger model of the dynamics of inquiry in which it is embedded. Then, I will provide an example from the history of science (John Snow's research on cholera) which illustrates the dynamics of inquiry. Finally, I will set out a detailed account of the functional complexity of evidence. I will conclude the first part by considering and responding to potential objections to the account.

In part two, I will argue that a variety of problems of evidence (listed above), which otherwise appear disparate, are unified by a common cause – the traditional model of evidence. It follows that what are often regarded as deep philosophical problems of evidence are, in fact, merely failures of a particular theory or model of evidence. By contrast, their resolution or dissolution seems almost trivial once we switch to the DEF-model.

Part I
The Functional Complexity of Scientific Evidence

2 Models of Inquiry and Evidence

2.1 The Traditional, Non-Dynamical Support Model

The default assumptions frequently relied on in discussions of evidence in philosophy of science lay out a family of theories or models of evidence which I will collectively refer to as "the traditional model." I will here briefly try to describe the main features of this problematic but common, often *implicit* model of evidence. An implicit model is an organized set of assumptions that plays a role in producing various kinds of judgments and reactions that nevertheless is not explicitly articulated or acknowledged by the one who relies on it.[7] In the case of the traditional model of evidence, it began life as an *explicit* philosophy or set of such philosophies, and continues to be relied on

[7] I'm using "implicit model" here in a way consistent with the use of "implicit theories" in psychology (Sternberg, 1985; Dweck et al., 1995) and the discussion of "nonexplicit philosophies" in (Drengson, 1982)(reprinted as Drengson, 2010).

to some degree because of the lack of a systematic alternative. Reliance on such a model seems inversely proportional to the degree to which one has attempted to grapple directly with providing a theory of evidence (part II provides examples of various ways in which those engaged instead in dealing with *particular* problems of scientific evidence fall back on assumptions of the traditional model).

In contrast to the model I will defend, the traditional model is:

(a') *Essentialist* – Evidence is defined by some *essential* property that suits it to stand as evidence.

(b') *Mono-functional* – Evidence plays *only* one important functional role: support for hypotheses or theories.

(c') *Non-dynamical* – Whatever the dynamics of scientific discovery might be, they are not relevant to understanding evidence. "Support" is an abstract, timeless relation between some set of evidence and some hypothesis.

(d') *Absolutist* – A bit of evidence is evidence regardless of context; anything that isn't fit to serve as evidence everywhere isn't fit to function as evidence anywhere.

I do not consider these four characteristics to be *necessary* conditions for membership amongst the traditional accounts that are contrary to DEF. Indeed, I hope to contrast my model to accounts that even hold weak versions of just one or two of these theses. But the more of these features a theory or model of evidence holds, the more problematic I regard it. Most central to the concerns of this paper are versions of (b') and (c'), but I include (a') and (d') for the sake of completeness.

Classical empiricism is a clear example of *essentialism* about evidence (a'): evidence is all and only impressions or sense-data which are immediately given and self-validating items of experience. Certain inductive logics provide another example when they require that evidence consist in *particular* propositions about (observed) matters-of-fact, while hypotheses are general propositions confirmed or falsified by such evidence. Most such accounts are also *absolutist* (d'), as are any accounts that require evidence to meet a non-contextual standard of certainty.

The traditional model is *non-dynamical* (c') in the sense that it doesn't depend in any important or interesting way on the *temporal complexity* of inquiry. This is not just a matter of historical context, but rather temporal structure. Traditional models of evidence may be temporal in the sense that the consider the belief available *at a time*, or that they take into account not only evidence and hypothesis but also *background beliefs* which are known to change over time. To carry through the physics analogy, these features might be said to constitute the *kinematics* of inquiry (its movement over time), whereas we're interested also in the *dynamics*, that is, structure behind the motion.

The traditional model is *mono-functional* (b') because it defines evidence according to a single function, the "support" relation it has to hypotheses, theories, claims, etc. As this is perhaps the central, most problematic, and most widely accepted feature of the model, one might easily call it the *support model*. Positivist and Popperian models from the middle of the twentieth century are clear specifications of the support model, as are some Bayesian accounts of evidence ("support" being understood as verification, falsifcation, or confirmation, respectively).

On the traditional account, "support" is an abstract relation that some set of evidence (beliefs, propositions, measurement records, etc.) holds to some further hypothesis or claim, whether the nature of that relation be logical, statistical, or formal in some other sense. Given a set of evidence and some hypothesis, we should be able to identify whether that set supports the hypothesis, and perhaps how much (at least well enough to rank-order hypotheses on the basis of the evidence). Further, we can always ask *at a time* what *the* evidence supports, and there is always a determinate fact of the matter (though we may not know what the answer is). The fact is not dynamically sensitive, i.e., sensitive to where we are in a process of scientific inquiry; it depends only on what the body of evidence is (and, perhaps, background beliefs). Evidence is that which justifies, and at a fundamental level it must be more certain, more justified, more secure than that which it justifies. That is, support is a one-way relation from evidence to hypothesis. Usually, evidence must also be independent of that which it justifies, lest the justification be illegitimate because circular.

While this may appear to be a caricature to some, in its basic outlines, this model captures the basic background framework for most contemporary discussions of evidence, despite explicit denials of one or more features. In Part II, we will see ways in which the traditional model exercises an *implicit* influence over important debates about evidence in science. You can also see explicit statements of commitment to aspects of the traditional model. While Eric Barnes' (2008) account of predictivism has a form of dynamism, he also assumes throughout that evidence is mono-functional, and so prediction is better than other evidence because it offers "stronger" support. (b') is almost ubiquitous, e.g., "Thus, for the Bayesian no less than for the Evidentialist, it is evidence which justifies that which stands in need of justification"(Kelly, 2008). According to Jim Bogen, "Much of the standard philosophical literature on... observational evidence tend to focus on epistemological questions about its role in theory testing," which is treated almost entirely as a matter of one-way support or justification relations (Bogen, 2010). Bogen and Woodward have argued that the role of observational data is not to support theories but rather to generate phenomena, whereas phenomena are used to support (or disconfirm) theories Bogen and Woodward (1988); Woodward (1989); Bogen and Woodward (1992, 2005). This introduces a distinction between data and phenomena as types of evidence, and a distinction between two types of functions for types of evidence: supporting theories and generating conclusions about phenomena (though in specifying the latter, they sometimes seem to collapse the distinction between these roles). This looks like a major step forward towards a more complex, dynamic framework, though the data-phenomena distinction has been quite controversial (e.g., Glymour, 2000). However, the role of phenomena in reaching conclusions in inquiry is pretty much the same as the traditional account of evidence.[8]

In the basic definition, the *Stanford Encyclopedia of Philosophy* entry on "Evidence" actually gets things right:

> Evidence, whatever else it is, is the kind of thing which can make a difference to what one is *justified* in believing or (what is often, but not always, taken to be the same thing) what it is *reasonable* for one to believe. (Kelly, 2008)

This is perfectly neutral between traditional and DEF accounts. The way in which evidence makes a difference to what one is justified in believing (or better, concluding, asserting, judging) does not have to be by way of a mono-functional, non-dynamical "support" relation, nor must we assume that evidence has any essential properties or that the relation of support is absolute. However, that same article frequently assumes that the way it makes a difference to justification is by way of such a relation. For example, consider the explanation of the total evidence condition:

> To the extent that what one is justified in believing depends upon one's evidence, what is relevant is the bearing of one's *total* evidence. Even if evidence E is sufficient to justify believing hypothesis H when considered in isolation, it does not follow that one who possesses evidence E is justified in believing H on its basis. For one might possess some additional evidence E', such that one is not justified in believing H given E and E'. In these circumstances, evidence E' defeats the justification for believing H that would be afforded by E in its absence. Thus, even if I am initially justified in believing that *your name is Fritz* on the basis of your testimony to that effect, the subsequent acquisition of evidence which suggests that you are a pathological liar tends to render this same belief unjustified. (Kelly, 2008)

Here it is clear that the author considers justification to be a one-way relation between a body of evidence and a hypothesis.

[8]See also Giere's (2006) account of model-testing, which, while adding some important layers, still comes down to a one-way, linear comparison between models of data and representational models similar to the traditional account.

Part of the problem is a lack of recognition of the *existence* of a model at work in philosophical discussions at all. It is quite easy to default to an ingrained model when one isn't aware of the existence of the model in the first place. Such models are the source of our claims about what is "obvious," "intuitive," or "almost true by definition" about evidence, but they are nonetheless revisable or replaceable.[9] Considerations of what seems obvious should bear little weight as compared to a theory of model that is descriptively and normatively fruitful.

2.2 Dynamical models

The temporal dynamics of inquiry have received insufficient attention among those interested in the nature of evidence. While it is popular nowadays to talk about science in terms of "practice," few have explored the impact that taking the praxical side of inquiry seriously for understanding the unfolding of science in time.[10] I am aware of only three detailed (types of) models of the temporal dynamics of science. One is the class of models developed by Kuhn (1996) and his followers (and here I include historicist critics of Kuhn, such as Laudan (1984, 1977) or Lakatos (1970), who provide different but related models at a similar scale (cf. Matheson, 2009)). This type of model discusses the career of large-scale theories, traditions, or research paradigms that govern entire disciplines or sub-disciplines over a large span of time. However, these models are so large-scale and long-term that they are not useful for addressing current concerns in the literature on the nature of evidence. By contrast, current issues deal *not* with the evolution of theories over the long run, nor the revolutionary replacement of theories or paradigms. The questions at issue – from the experimenter's regress to contemporary concerns about the role of evidence in policy – are far more local than these accounts can address, having to do with with the role of evidence in single controversies within a discipline or paradigm. To put it differently, the theories of Kuhn and Lakatos are concerned with the dynamics of *theory-change*, not the dynamics of *inquiry* (where there may be no theory change).

One point that is central to Lakatosian philosophy of science which has also received some attention in more traditional confirmation theory is the idea that *novel prediction* is particularly important, that it is what matters

[9] This is one of the great contributions to philosophy of John Dewey and Richard Rorty, to show that philosophy, like science, gets at the world through sophisticated but optional and replaceable *theories* or *models*, and that often what we need is not to answer certain questions or solve certain problems but to replace the theory in which that question or problem is stated. An importantly related idea is that of "metaincommensurability," discussed by Oberheim and Hoyningen-Huene (1997).

[10] Wayne Martin comes close in *Theories of Judgment* (2006) when he argues that the temporal complexity of *judgment* has been ignored, though in the end he has little specific to say about what this temporal complexity looks like.

most of that it has a certain special status. Clearly, if prediction is what matters, and prediction is always prospective (there are non-temporal accounts of novel prediction, of course), then evidence depends on a certain kind of dynamic relation between hypothesis and evidence. It is telling, in terms of the hold of the traditional model, how many philosophers have found great difficulty explaining the importance of prediction, or have tried to reduce the dynamical quality of prediction to standard non-dynamical approaches. However, the predictivist account threatens to reduce its dynamic complexity if it downplays too much the role of prior observation evidence. Likewise, some versions of predictivism assimilate the function of predictive evidence to "support." In such accounts, predictive evidence simply lends *more*, *stronger*, or *better* support (see Barnes, 2008, p. 1). More sophisticated accounts of predictivism may have more in common with the DEF-model than the traditional model.

Another dynamical model of inquiry is the pragmatist model introduced by Charles S. Peirce and further articulated by John Dewey.[11] This model works best at the more local level of particular scientific inquiries, though it has some applications at the larger scale.[12] In Peirce's original formulation, *doubt* is a necessary condition for genuine inquiry of any sort, the sort of doubt that arises when previously held beliefs and habits of action[13] fail to guide one through a particular circumstance. Inquiry, then, is the process of responding to doubt in order to fix new beliefs and habits that resolve the doubt and allow activity to continue. The temporal structure of inquiry depends on this movement from uncertainty through investigation to settled belief. Dewey adopts this basic structure,[14] supplementing it with an account of the internal complexity of inquiry, the phases of reciprocal adjustment between fact-gathering, hypothesis-forming, and experimental testing that lead to what Dewey calls "warranted assertion" or "judgment" rather than merely "belief."

2.3 Functionalist Theories of Evidence

Essentialism and absolutism are the aspects of the support model whose fortunes have been the worst (as mentioned above, the parts of the model need not always go together); both have been explicitly denied in various ways, and find few defenders amongst contemporary philosophers of science (though their fortunes have been fairer amongst ordinary epistemologists). Thus, I will focus primarily on defending (b) and (c) over (b') and (c').

To deny essentialism is, in my terms, to assent to *functionalism* about

[11]See Peirce (1877); Browning (1994); Dewey (1938); Hickman (1998).

[12]I am not implying that there are important conflicts between Kuhnian models of scientific development or predictivism and the pragmatist theory of inquiry.

[13]This formulation is redundant if we adopt Peirce's definition of belief.

[14]In Dewey's terms, inquiry is a transformation of an indeterminate/problematic situation into one that is settled.

evidence. Functionalism is most familiar from philosophy of mind, where it is the view that what a certain kind of mental state (e.g., a belief) consists in is *not* dependent on its constitution (e.g., an idea in my spirit-substance or a configuration of neurons in my brain), but rather on the *role* it plays in my cognitive economy, most simply conceived as its causal relationships between perceptual inputs, behavioral outputs, and other mental states.[15] For example, a belief may be caused by certain perceptual inputs and inferential operations performed upon them and on other beliefs, it may have causal relationships with other beliefs, and, when combined with desires, it may cause certain behaviors. The sum of these relationships is the *functional profile* of a belief, and, if functionalism is true, then that profile is *all it is* to be a belief. As far as its constitution, that belief could be anything including non-extended mind-stuff, a configuration of neurons, or the circuitry of a suitably complex artificial intelligence. Likewise, a certain collection of neurons might well change from belief to something else if its functional role in the mechanism changes over time.

In its basic form Bayesian epistemology is a form of functionalism about evidence. For Bayesians, evidence is the E that figures in formulae like $P(H|E)$ (conditional probability of hypothesis H given E), $P(E|H)$ (likelihood of H on E), etc., and used to conditionalize beliefs, calculate degrees of confirmation/disconfirmation, etc.[16] For all practical purposes, this is *all it is* to be evidence for Bayesianism. Often, it is implicitly or explicitly stated that E must be a *statement*; however, nothing in the basic theory requires this. It is just as reasonable to suppose that a telescopic image or the results of a computer simulation can function as evidence, so long as you can assign the needed probabilities to it. Likewise, even when considering statements, nothing requires that our evidence be a *particular* statement, or one referring to *observational facts*. Unless one adds restrictions to the contrary, anything that gives a conditional probability for H can serve as evidence, even something more general that H itself. Bayesianism even has a primitive sort of dynamism in that it requires one to update one's degrees of beliefs on acquiring new evidence; however, it is not at all clear what sort of events in actual, concrete scientific practice instantiate this abstract operation.[17]

---

[15] See Levin (2009) for an overview. Needless to say, the comparison to functionalism in the philosophy of mind is merely an analogy, to demonstrate the functionalist style of explanation. Nothing in my account hangs on the success or failure of functionalism about mental states. Functionalist theories have also been given for the ontology of colors (Cohen, 2009), truth (Lynch, 2000, 2001; Wright, 2005; Lynch, 2005), and morality (Jackson and Pettit, 1995, 1996), whose fortunes are likewise independent from my account.

[16] See Talbott (2008) for an overview.

[17] Possible Answer: Inference.

Response: It is not at all clear what sort of events in actual, concrete scientific practice instantiate "inference" in this (abstract) sense.

...

2.4 Other Functions Beside Support

While Bayesianism is an example of a functionalist theory of evidence (and thus an advance, on my view), it is a rather simple, impoverished one. To see this, we can return to the analogy with philosophy of mind. The simplest version of a functionalist theory of mind (so simple as often to be regarded as a *precursor* theory to functionalism proper) is *philosophical behaviorism*. On that view, the functional profiles of a mental state are specified exclusively in terms of the relationships between perceptual inputs and behavioral outputs (i.e., no causal relationships are allowed *between* mental states). So, to believe that the earth is round or to feel angry just is to respond with the right behavior given some stimulus. This sort of theory of mind is now widely regarded as too impoverished to do its job, i.e., to account for what mental states are. Functionalists argue that this is because behaviorism ignores the relationships between mental states.

Likewise, I will argue that the traditional model, even functionalist versions like Bayesianism, is too impoverished to do the job. In this case, it fails to provide a theory of evidence that fully accounts for the ways that evidence functions to bring an inquiry to successful resolution. It is too impoverished because it only allows for a *single* functional role for evidence, the role of supporting a hypothesis (theory, claim, etc.). By contrast, as I will argue in the rest of the paper, I think we can point to a number of equally essential roles that evidence plays in inquiry.

This point is common amongst philosophers of scientific experiment. As Ian Hacking has said,

> Experiments, the philosophers say, are of value only when they test theory. . . So we lack even a terminology to describe the many varied roles of experiment. (Hacking, 1982, p. 71)

And in a similar vein, Allan Franklin has argued that

> Experiment plays many roles in science. One of its important roles is to test theories and provide the basis for scientific knowledge. It can also call for a new theory. . . Experiment can provide hints about the structure or mathematical form of a theory, and it can provide evidence for the existence of the entities involved in our theory. . . it may also have a life of its own, independent of theory: Scientists may investigate a phenomenon just because it looks interesting. Such experiments may provide evidence for future theories to explain.

My account goes further by enumerating the various roles of evidence (observational *and* experimental) and showing how they fit together to guide inquiry to successful conclusion.

3 The Dynamics of Inquiry

I will now begin to outline systematically the DEF-model of evidence as an alternative that is truer to the complexities of scientific inquiry and avoids the vicious simplifications of the support model. To begin, I will give a description of the dynamics of inquiry in which, according to the DEF-model, evidence is embedded. The purpose of this section is to describe the functionalist model of the dynamics of inquiry, including the complex functional roles for evidence within that process. The following section illustrates the fruitfulness of the model by showing how it applies to an interesting case of scientific inquiry, John Snow's work on cholera.[18] After summarizing the main roles that evidence plays in inquiry, I entertain some basic objections to the framework and offer replies. In Part II, I will show the power of the DEF-model in solving (or dissolving) some major problems in the philosophy of scientific evidence – this, I think, is the most compelling argument for the view. At this level, it seems to me, there is no way to give more direct arguments for the view that aren't also question-begging.

In the main outlines, the dynamics of inquiry[19] can be described by a number of interlocking phases (see Figure 1):

1. *Inquiry begins with a felt perplexity.* There are many types of perplexity, but they are not in general a mere state of ignorance on the part of the inquirer. Rather, the objective state of the science—which may include theoretical frameworks and concrete models, techniques of observation and sets of data, methods of prediction and expectations of inquirers, and so on—is contradictory, confused, indeterminate, or in tension. There are conflicting tendencies within the situation of the field at the present time, a major discoordination of the practice, and this requires investigation. Hence, there are affective, practical, and objective aspects of the perplexity or indeterminacy. (Contrast perplexity with the smooth application of some theory or technique to a case with immediate success.)

2. *Discrimination. Operations of observation* must take place in order to take stock of the situation that evokes inquiry. We need to gather data on the situation that helps us begin to understand the problem at hand and the

[18] It is ultimately an illustration only. It is obviously absurd to suggest that one could defend an account of this kind on the grounds of some sort of induction on a single such (or any small number of such) cases.

[19] This model is loosely inspired by John Dewey's version of the pragmatist theory of inquiry. It is not, so far as I can see, committed to any of the more controversial pragmatist claims about truth or meaning. If the following sounds a bit like the description of "The Scientific Method" from an elementary science textbook, don't be too surprised: Dewey was influential over the shape of science education, especially in America, though his ideas have been vulgarized. Careful scholars should not consider the association a black mark against Dewey's views.

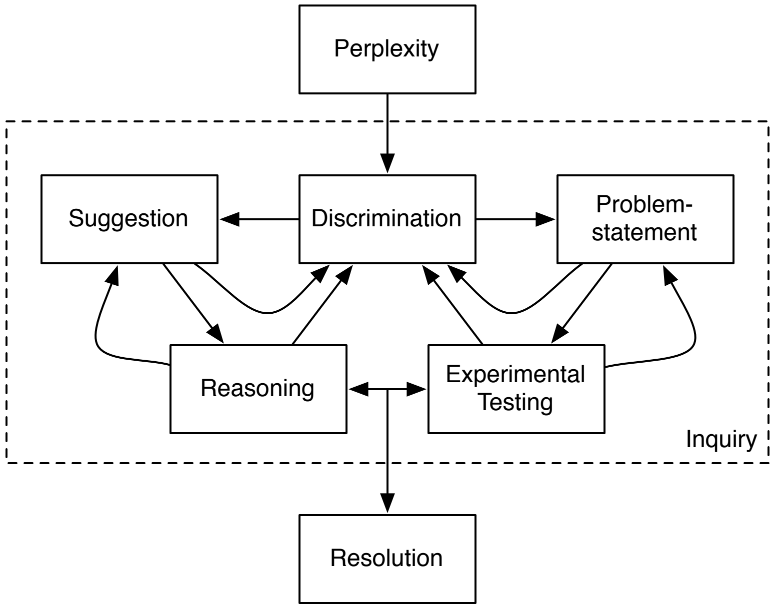

Figure 1: Boxology of the **functional dynamics of inquiry**. (Connections between phases have been simplified for clarity.)

conflicting tendencies in our response to it. Prior to the interruption that begins the inquiry, the distinction between conceptual and observational materials is vague. In habitual activity, we tend to run together the facts and our ideas about them, and we behave as if there is no difference between the model and the thing. This is a reasonable and necessary way to go on, so long as no problems arise. But problem-solving inquiry requires that we discriminate (a) the factual vs. conceptual materials we have to work with, (b) features of the subject-matter in question. These constitute the relevant features of the situation which has become perplexed, and are required to determine the nature of the problem and our response.

3. *The Statement of the Problem.* The situation must be assessed in order attempt to *formulate a problem-statement* that adequately captures the given perplexity. Scientific inquiry does *not* begin with a set problem or question at which science is directed. The agenda of inquiry cannot be set by fiat. Where no genuine perplexity exists, there is no room for scientific inquiry. Where it does, the problem cannot be accurately or adequately stated ahead of time; the statement of the problem is a phase of the inquiry itself, and it evolves as the inquiry is pursued and more adequate and sophisticated observations are made.

4. *Suggestion of Hypotheses.* The first pass at determining the factual conditions of the situation, the conceptual possibilities in our theories, and the terms of the problem *suggests hypotheses* for solving the problem. Forming

a problem-statement and suggesting a hypothesis are coordinate activities. The former connects to the settled features of the situation in which a tension arises, while the latter connects to some possibility for further action that resolves the tension. If the factual side of inquiry pertains to what has been determined, then the hypothetical (conceptual, theoretical) side of inquiry pertains to what is possible. (This is the process that theories of *abductive reasoning* are trying to analyze.)

5. *Reasoning.* A reciprocal process of *coordination* of observed facts and theoretical-hypothetical ideas is undertaken. There are several aspects of this process which depend on each other and need not proceed linearly.

    a. Background theoretical materials, well-tested models, and other conceptual resources are brought to bear on the problem at hand.

    b. Hypotheses are developed by processes of reasoning to be more specific and relevant to the case at hand, to be in greater concert with more general theoretical materials, to suggest further operations of observation, and to take into account the evolving body of data and statement of the problem.

    c. New observations are made in response to the evolving series of hypotheses and theoretical ideas, to answer questions posed by them and fill in information needed to specify the relevant features of the ideas.

    d. From the set of putative evidence constructed so far, certain are selected or amplified as *relevant*, while others are rejected as irrelevant, imprecise, poorly executed, or explained away as effects of interfering phenomena that must be controlled.

    e. The statement of the problem is refined to reflect the changing understanding of the situation and the evolving series of hypotheses.

6. *Experimental testing.* A series of controlled, limited, or tentative, *experimental applications* of the hypotheses are made in order to evaluate their probable efficacy in solving the problem. Earlier experiments can suggest more refined experiments, or the necessity of further articulating data and hypothesis, or the need to "go back to the drawing board."

7. *Resolution.* The aim and final product of inquiry is a judgment of how to proceed, how to resolve the perplexity that initiated inquiry. Inquiry continues until one of the hypotheses is adjudged to be the most warranted amongst the alternatives, and the alternatives have been more or less ruled out. To put it differently and more prospectively, the inquiry proceeds until a point of resolution so settled that the conclusion can be used as a reliable means to further inquiries. A judgment of warrant is a judgment about the adequacy of the hypothesis to solving the problem. Such a judgment

> is impossible without to some degree undergoing this process of inquiry (otherwise, it would be merely a reflexive response), and ideally the process of inquiry must be exhausted to the point that no doubt remains about the hypothesis, and the conflicting tendencies of the situation have been resolved and coordination has been restored (at least, for the moment, for the most part).

This is obviously an *idealized* picture of the conduct of inquiry.[20] It is no *a priori* imposition, however; it is informed by reflection on the complexities of the history of science and scientific practice.[21] It is a *normative-explanatory* model, attempting to capture, explain, and make available the lessons of successful inquiries past, as well as incorporating general cognitive and epistemic considerations. The proof of this model is in its power to give us a more successful understanding of the uses of evidence and to resolve or dissolve problems of evidence that arise. If the account seems overly simple, all to the better; my main point is that almost all philosophers working on these problems are using an even simpler model, and I would be happy to entertain even more complex alternatives. I do believe that we have to make some sacrifices in the direction of simplification in order to have a usable, systematic framework, and this too has guided my focus.

Next, we will look at an illuminating concrete example of inquiry, in order to show what lessons for understanding evidence this model provides. Throughout the next section, I will indicate the interplay of the different phases of inquiry.

4 Snow on Cholera

Consider the work of John Snow on the transmission of cholera.[22] The basic outlines of the *perplexity* (**Phase 1**) are clear: cholera is a terrible disease, fatal in nearly all cases at the time. The nineteenth century saw many epidemics of the disease, beginning in Asia and later in Europe and America. It is

[20]It is also worth pointing out, I think, that not only is there plenty of inquiry that isn't particularly concerned with high-level theories, but also there are activities in science that do not constitute problem-solving inquiry at all – those involving education, training, exploratory "problem-finding" research, to name a few. Both of these insights are tied up with the experimentalist slogan, "Experiments have a life of their own." I have little to say about the latter set of activities, except to say that they are not primarily evidence-gathering activities, except retrospectively insofar as they turn out to spur inquiry.

[21]I do not intend to suggest that it is reflection on the particular case in §4 that justifies the model. Rather, it derives from the work of Peirce, James, and especially Dewey – all first-rate scientists in their own rights, from my own personal experience, and from reflecting on history, sociology, and empirically-sensitive philosophy of science. I also mean to suggest that the above account ought to be subject to empirical critique.

[22] My discussion here is taken from Goldstein and Goldstein (1978, 25–62) who draw heavily on Snow's own manuscripts. Parenthetical page references are to their discussion.

tempting to say that the *problem* (**Phase 3**) itself is clear from the beginning: how is cholera communicated, and how can its transmission be prevented or contained? While the idea of contagious diseases was not new in the middle of the nineteenth century, when Snow was at work on cholera, it was neither fully accepted nor clearly distinguished from views identifying disease as a punishment for sin. To regard some diseases as communicable, and to identify cholera as one such, is already to be well into the inquiry. Understanding the exact nature of the problem is especially difficult because the transmission of cholera didn't follow the expected pattern of the prominent "effluvia" theory of contagion, according to which disease was transmitted by emanations or exhalations from the sick patient into the surrounding air. Cholera tended to be concentrated amongst the poor, and almost never infected the doctors who tended to the sick. This was taken as evidence that the disease was "a just punishment for the undeserving and vicious classes of society"(26). To regard the problem as one fixed prior to inquiry would be to *falsely* take as antecedently determined many things that are at first unsettled.

Snow begins by collecting a variety of general and fairly pedestrian *facts* (**Phase 2**) (p. 29):

1. Cholera began fairly localized in India, where Europeans first encountered it in the late eighteenth century and spread rapidly from there in the early nineteenth century.

2. Cholera travels along channels of and at the speed of human interaction, always appearing first at the sea-ports of new islands and continents, and it never attacks those sailing from countries free of the disease until they enter the port or come ashore in a place where the disease is found.

He then moves to more specific cases (30–1):

3. Mrs. Gore's son, who had been living and working at Chelsea, came home ill and died of the disease. His mother, who attended him, caught ill the day following his death, and was dead the day after. No other deaths from cholera in the area took place.

4. John Barnes died after having contact with the clothes of his sister who had died from the disease, whose personal effects had been sent to him after her death.

5. Mrs. Barnes, who contacted the illness from her husband, was attended by her mother, who contracted the disease after washing her daughter's linen.

And so on. All of this clearly *suggests* the idea (**Phase 4**) that *the disease is communicable.* Further, it might naturally suggest the most common explanation of the transmission of disease, *the "effluvia" theory* already mentioned.

Already the cases of John and Mrs. Barnes suggests some difficulty with this explanation, since John Barnes was never exposed to his sick sister, and Mrs. Barnes' mother was healthy while she was in the presence of her daughter, only contracting the disease after contact with her linens. Here, Snow has begun *reasoning* about the facts and hypotheses on offer (**Phase 5**). Further evidence tells against this hypothesis (31):

6. It is not always the case that someone who spent time in the same room with the patient, or attending to them, is likely to contract the disease.

7. One need not ever come near to the patient to contract the disease.

8. Other diseases such as "the itch," syphilis, and intestinal worms are transmitted by vectors other than air.

9. The pathology of the disease begins with intestinal symptoms, rather than any symptoms of systematic infection such as fever.

The final two pieces of evidence suggest another hypothesis: *The disease spreads by some infected matter "ejected" from a cholera patient being accidently ingested in sufficient quantity, and whenever this accidental consumption of infected matter is likely, the disease is highly likely to be communicated* (33).

This hypothesis suggests some further observations. If it is valid, you'll find that certain people who come near to the patient do not get cholera (as we've seen), and further that they avoided it by way of habits of cleanliness that would prevent them from accidently ingesting any cholera evacuations. Indeed, this is clearly the case with doctors:

10. Doctors do not generally contract cholera from their patients, while persons who attend to the patient in a more personal way, with less concern for cleanliness, are more likely to contract the disease. (33)

Reasoning through the implications of the hypothesis (**Phase 5**), we can see that there are several reasons that people of different social classes would have different risk of contracting the disease: they "perform different functions around the sick", live in different conditions, have different lifestyles and personal habits concerning cleanliness and quantity of human contact (33). Further general observation tends to bear out the hypothesis (34). One bit of evidence raises a puzzle, however:

11. Cholera does sometimes spread to the rich despite the absence of the vectors of direct communication present in the case of poor laborers.

In other words, rich folk live in much less cramped environments, tending not to "live, sleep, cook, eat, and wash" in the same space (34). They do not usually tend intimately to sick persons, or if they do so, they wash carefully and constantly. It seems very unlikely that the illness would spread between family members in such circumstances, and it rarely does.

Nevertheless, rich people *do* contract the disease in sizable numbers in some cases. Snow did not take this to invalidate the hypothesis, however. Rather, he supposed a further specification of the hypothesis in these cases that would provide the appropriate kind of transmission vector: *cholera can spread through the water supply* (35), and further cases support this hypothesis.[23]

Having worked out the implications of the hypothesis and found corresponding facts is not, however, where Snow stopped. At this point, his hypothesis is surely plausible, but not firmly established. The next phase requires *experimental application* (**Phase 6**) of the hypothesis to real situations in order to test its adequacy. This goes beyond merely collecting observations about cases of cholera, either individually or in bulk. Experiment is *not*, as many have supposed, just a special way of generating further observations. In many ways, and in many cases, the procedures may look very similar. Certainly, techniques of observation are part of the experiment, and experiments may even produce data that are fed into the recursive process of coordinating facts and ideas, but the *functional role* is nonetheless very different. The roles of *observation* are to fix the conditions of the problematic situation and the terms of the problem, as well as to suggest and refine hypotheses. The roles of an *experiment* is to put the hypothesis *into practice*, in a limited and controlled fashion, in order to determine its efficacy in solving the problem. It is not so much the nature of the process producing the evidence that determines its type, then, as the role or function of that evidence in inquiry.

Snow engaged in at least two experiments, neither of which was entirely satisfactory from the point of view of the model under consideration. His first experiment was with the Broad Street water pump in 1849. In this case, by first observing the circumstances of a certain outbreak of cholera, he was able to determine, based on his hypothesis, the probable cause of the outbreak in the pump on Broad Street. He was able to determine that use of water from that pump was a common cause of most cases of the outbreak (37–39). Likewise, he was able to determine that amongst the groups in the area who were mostly *unaffected* by the outbreak, all had avoided, for one reason or another, use of the pump (39–40). He then made an experimental intervention by convincing government officials to remove the handle from the pump to prevent its use. Unfortunately, removing the pump-handle failed to produce any significant effect on the number of new cases, and this is likely because the epidemic had pretty much subsided by the time of the experiment. So, while

[23] It isn't clear from Snow's reports what the order of inquiry was supposed to be at this point between finding cases and framing the hypothesis.

there was plenty of supporting evidence for the pump as cause of the outbreak (including indirect evidence of the contamination of the water by sewage), the experiment failed to be conclusive (40–41), because the intervention failed to have any appreciable effect on resolving the problem due to the fact that cases of cholera were already in rapid decline for other reasons.

Snow's second experiment was what is sometimes called a "natural experiment" (42). There were no actively controlled circumstances, nor were there even any active interventions. Instead, a "natural experiment" is one in which the natural course of events is such as to be *as if* one had set up an experiment to test the results. In the London cholera outbreak of 1853–4, Snow was able to find a very distinctive pattern in the deaths resulting from cholera according to which of two water companies in London supplied the house with water. Snow's study had two parts. First, using what we would today call "retrospective study design," Snow began with a district of London in which houses were supplied by two different water companies—Southwark & Vauxhall or Lambeth—in fairly random mixture. He then looked at all of the reported cases of cholera in that district, determining that of the 44 deaths, 38 were suppled by the Southwark & Vauxhall Company. In the second study, using what we would today probably call a "prospective design," Snow looked at all of London by water company, and discovered that the rate of deaths from cholera in houses supplied by the Southwark & Vauxhall Company was an order of magnitude larger than either those supplied by the Lambeth Company, or among houses supplied by neither (some third party, local well, etc.). Snow argued that the connection between houses and water companies was quite randomly distributed with respect to the relevant factors (two neighbors were even in some cases supplied by the two different companies). This affords a significant test of the hypothesis: it seems difficult to deny that water supply has in this case had a significant effect on incidents of the disease, or that the "act" of avoiding the contaminated water supply significantly reduced the risk of contracting cholera (42–46).

While Snow performed no active interventions in this case, it still *functions* as an experiment, plays an experimental *role.* It is not any particular technique that makes the experiment, and Snow need not even have engaged in any direct intervention. What matters is the function it performs, the way that the experiment is *taken up* in the process of inquiry: as an application of the hypothesis to the situation. Nevertheless, one would prefer a more active application of the hypothesis to the problem of cholera, based on the model I've described, because this would serve more directly as a test of the ability of the hypothesis to act as a problem-solution, to move the problematic situation towards resolution. A careful analysis of past events can be of very significant use in the course of an inquiry, especially in the rare case where we have sufficient information to analyze the events *as if* they were a deliberate experimental intervention. But ultimately, experimental inquiry looks forward, towards a transformation of the situation from problematic to settled, rather than backward at what has come before. Hence, when possible, we prefer an active intervention that changes present conditions.

Though he offered further evidence for his hypothesis, Snow never produced such a test. He did provide further support for his theory, however. He rejected certain apparent counter-evidence by providing reasons to regard it as either irrelevant to, or explicable in a way that was compatible with, the main hypothesis. He combined reasoning and observational evidence to provide arguments for rejecting alternative hypotheses. And he described analogous suggestions for other diseases, whose causes were both known and unknown. All of these fit well within the model above, under heading (5): the reciprocal coordination of factual and hypothetical materials through reasoning. Snow uses observations to help select and refine a hypothesis, and he uses a guiding hypothesis to discriminate putative data, in a reciprocal process that arrives at a tight fit between fact and hypothesis.

The final part of Snow's monograph on cholera is the most crucial, from the point of view of our model, though I suspect it has rarely been regarded as so by other commentators.[24] In the last section, Snow provides a list of twelve recommendations for how to prevent the spread of cholera, based on his two hypotheses, plus some further reasoning about possible cases. For example:

> 1st. The strictest cleanliness should be observed by those about the sick...
> 3rd. Care should be taken that the water employed for drinking and preparing food... is not contaminated with the contents of cesspools, house-drains, or sewers; or, in the event that water free from suspicion cannot be obtained, it should be well boiled...
> 11th. To inculcate habits of personal and domestic cleanliness among the people everywhere... (53–4)

And so on. In the long run, it seems to me that these recommendations are crucial to the eventual acceptance of Snow's explanation, and hence, the ultimate resolution of the inquiry (**Phase 7**). The fact that the problem was resolved as far as Snow himself was concerned is relevant to inquiry from a purely personal point-of-view, but for a truly scientific inquiry, social dissemination and understanding, according to the inquiry-model, are crucial to the final judgment of the hypothesis. Further, no amount of convincing argument or "decisive proof" provided by a scientific manuscript can be the ultimate measure of a scientific judgment. Scientific claims must be judged, on the one hand, by others taking the results to be so settled as to provide a steady resource for further inquiry and, on the other hand, by the success of future

[24] Goldstein and Goldstein, for example, include it in a section near the end of their paper entitled "Applications to Other Problems" (51ff), and treat it as something of an afterthought.

applications, such as the ones suggested by Snow in this final section. It is the success of these further applications that are the "decisive experiments" that justify Snow's view, rather than any proofs that Snow had produced.

5 Evidence on the Inquiry-Model

Having laid out and explained the functional dynamics of inquiry, I can now set out the basic picture of the DEF model. (See Figure 2.) First, in the model of inquiry I've been discussing, *functionalism* guarantees that many different types of things count as evidence: not only particular, observed facts, but also historical developments, statistical analyses, general trends, "phenomenological" laws, and anything that adequately serves some part of the functional roles of evidence and some stage of the inquiry. Second, it is important to notice the very different roles that evidence plays in the course of an inquiry. In many contemporary accounts, evidence is, if not mono-modal (or essentialist), at least mono-functional: all evidence serves as a test of a theory or hypothesis, and it confirms or disconfirms it, or renders it more or less plausible, probable, or credible. On my account, evidence is not only multi-modal, but serves a variety of purposes (parenthetical numbers here refer to items from the case study in the previous section):

I. *Observational evidence* serves a variety of roles related to the way that operations of inquiry depend on an understanding of the present conditions that have led to some perplexity.

   A. Through *discrimination*, it provides information about the conditions of the problematic situation (3–5).

   B. It helps locate and state the *problem* (1–2).

   C. It *guides* speculation and hypothesis-formation (3–5)

   D. It guides *reasoning* in order to helps *eliminate*, *specify*, *clarify*, or *improve* our original hypotheses (6–11).

II. *Experimental evidence* serves the additional role of

   A. Tentative application of a developed hypothesis to check its consequences for future action and inference (the Broad Street pump experiment and the study of water companies).

   B. Generation of further observational evidence (generally of a very precise but specialized nature).

Experimental evidence in this sense, again, can be of many different kinds: not just controlled manipulations in a laboratory, but also "natural experiments" that function *as if* there were a manipulation (as in Snow's water companies

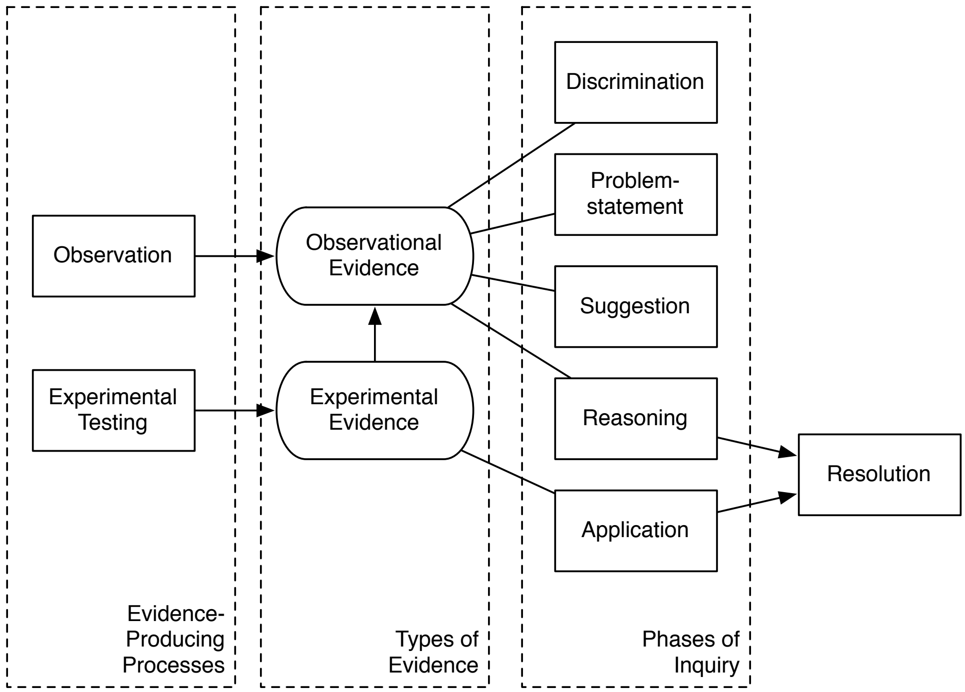

Figure 2: Boxology of the **Dynamic Evidential Functionalism** model.

research), as well as cases such as a change in public policy in a particular area whose consequences are then tracked to determine whether the application is successful. This is because it is the *functional role*, not details about the production of the evidence, that determines which evidence is experimental.

In every case, it is not some abstract or formal relation between the evidence and the hypothesis by which the evidence serves to justify the hypothesis. The formal and symbolic is only one side of evidence. It is rather a very concrete process of transforming a perplexity into a resolution that evidence is instrumental towards, and which ultimately justifies any final judgment of the inquiry.

This model has several benefits. First, it is more faithful to the complexities of scientific practice, in that it refuses to reduce the (philosophically relevant) activities of science to judgments of which hypothesis is best supported by the body of evidence, that it makes clear the ways in which data gathering is directed towards various ends, and that it reduces some of the mystery in the process of hypothesis-generation by proposing that hypotheses latch on to possibilities suggested by the facts of a particular situation. Second, it retains and strengthens the philosophical attempt to explain why scientific methods work, by describing the phases of scientific inquiry and how they work together in problem-solving. Third, it provides the strongest way of responding to the various problems of the "empirical basis" (e.g., epistemic status of evidence, theory-ladenness, experimenter's regress), which will be the focus of Part II.

This more complex model of the functions of evidence can be used for a multi-scale analysis of the *functional fitness* of evidence, which gives as a way of assessing the adequacy of it to stand *as* evidence.

How can we be certain that some body of putative evidence *is* evidence? For traditional empiricist accounts, the answer appeals to the incorrigible and indubitable nature of particular sense-data. In contemporary accounts, the assumption is usually that evidence has a high degree of credence relative to our initial credence in hypotheses. On the DEF account, putative evidence and suggested hypotheses are both judged by their ability to be able to brought into mutual coordination, leading to a solution of the original problem. Many "facts" may be collected along the way, may aid in various functions in the course of the inquiry, but may be eventually discarded as being inadequate and replaced by new facts. At the end of an inquiry, the inquirer produces a chain of reasoning from general considerations to a specific hypothesis, as well as a body of evidence in support of that hypothesis. The chain of reasoning does not represent the actual steps in the inquiry that produced them, nor does the body of evidence include every bit of data gathered along the way. They are as much the conclusion of the process as the final judgment, and they are what we see reflected in ordinary scientific articles. That these final products cohere is essential, but mere coherence is insufficient: they must also cooperate to resolve the perplexity which spurred the inquiry. This is non-trivial because, as you will recall, the *perplexity* is not merely verbal or intellectual, but has affective, practical, and objective elements. Real re-coordination must be achieved. Evidence functions in the complex and dynamic ways laid out above to move an inquiry towards resolution; the evidence itself is thus evaluated in terms of its functional fitness in the process aimed at doing so.

6 Preliminary Objections and Replies

It is important to recall that the model at hand is an idealization in several senses. It is idealized in that it is *simplified*: it does not even pretend to capture every important element of scientific practice. It is nonetheless a useful idealization: the clarity it lends to particular cases such as Snow's, and more importantly the ease with which it resolves or dissolves a variety of puzzles about the the nature of evidence that plague contemporary discussion, will demonstrate its usefulness. Furthermore, it is less simple than every other model of evidence available, so any criticism of its simplicity will apply much more so to the various traditional approaches explicitly or implicitly critiqued in this essay.[25] It is also an *ideal* model, that is, it makes some modest *normative* claims. It hopes to capture something of the lesson of *successful* inquiries

[25]In other words, it is a criticism I would gladly accept, and I would encourage the other to proposed a less simplified but still useful alternative.

of the past. The model is about the best (the ideal) way to carry out inquiry. It is ultimately an interpretive model: individuating inquiries is a tool of the inquirer into inquiry (i.e., the philosopher of science, epistemologist, or logician), and the divisions need not be clear within primary inquiry, to the first-order inquirers themselves.

Likewise, it is crucial to point out that the model makes no claim that science is generally or usually a *large-scale* movement from less to more certainty or from more to fewer open problems – it is unlike Kuhnian, Lakatosian, etc. models of science in not making such large-scale claims. The ubiquity of open problems in scientific research suggests otherwise, certainly. Perplexities arise in many ways: from failed application, from new evidence garnered elsewhere, from contradictions between otherwise well-confirmed and highly general theories, from worries about the form or aesthetic qualities of theories, as by-products of other inquiries, and so on. Scientists positively go hunting for problems to work on;[26] by searching for potential problems, they secure in advance new ways of coping with the world and stabilizing practices that could otherwise become unsettled in tragic fashion. Nevertheless, something like what is described in the model in question, it is claimed, goes on once they fix on some perplexity and set to work on it in a fashion that tends to lead to success.

One might respond to the DEF-model that all the other supposed functions of evidence are really just consequences of the support function; thus, the DEF-model just reduces to the traditional mono-functional model. For example, one might argue that some evidence can only "suggest" a hypothesis by supporting it. Likewise, evidence can only guide problem-formulation by acting as support for a hypothesis of the form, "Problem-statement P accurately describes what is problematic about this situation." In a similar vein, we might keep "support" separate from "suggest" by defining the latter in terms of "taken to support," i.e., to say that the evidence E suggests hypothesis H, what we mean is that, for now, we take E to support H.

The key question we should ask ourselves in the face of this objection is what the purpose or value of this type of reconstruction is. Does the objection insist that, either cognitively or sociologically speaking, the operations of formulating problem-statements, suggesting hypotheses, etc. use precisely the *same* mechanisms as deciding which hypothesis best fits a body of evidence? This seems highly unlikely. Is the goal to justify all of these functions in terms of some *a priori* account of what inquiry *must* look like or what evidence *must* be? On the contrary, I insist we should look instead at the actual practice of successful inquiry, rather than our philosophical intuitions, to learn what evidence and justification look like.

In other words, the attempt to explain all of the various essential functions in terms of the support-function, taken to be the only *fundamental* function of

[26]This point is made in the context of Peirce's theory of inquiry by Browning (1994).

evidence, is a distortion of the phenomena in question. However the process of "suggestion" works (a fruitful area for psychological and philosophical investigation), I see no evidence that it is by a process of enumerating candidate hypotheses and determining which is best supported. Rather, suggestion is the process of generating those hypotheses in the first place. Likewise, the initial attempt to gather evidence in order to take stock of the situation and determine the nature of the problem at hand, to sort the ordinary and familiar aspects of the situation from the puzzling and problematic ones, doesn't look like assigning likelihoods or credences to competing propositions of the form "The problem here is P." One important reason for this is that an adequate problem-statement has to be *much more* than true; accurate but unfruitful problem-statements are no more helpful than inaccurate ones. The different functions of evidence are subject to epistemic values orthogonal to considerations of mere warrant or truth.

The attempt to treat evidence and support apart from the rest of the process of inquiry has been a dead end. My proposal is an attempt to get around that: if we work out a more complex functional profile for evidence, and more generally, a more complex picture of the varied interactions and relations facts, data, and evidence have with theories, models, and hypotheses, we will find a more illuminating picture. Of course, the proof of the pudding is in the eating, and I've merely proposed a new recipe. Whether proves fruitful depends in part on work that will be presented in Part II.

One might also respond to the DEF-model by drawing a sharp distinction between epistemic and causal, pointing out that *support* is epistemic and non-causal, while the other "functions" I've enumerated, like *suggestion*, are merely causal and non-epistemic. This explains why *support* is atemporal, non-contextual, etc., while the DEF-model is dynamical and contextual. Far from being a virtue, this is an indication of how my account conflates an important distinction. Many things may *cause* me to generate a hypothesis, including a hallucination or a knock on the head, but such things are not evidence. Evidence is just what supports a hypothesis. In other words, insofar as suggestion is independent of support, it is purely a causal relationship.

I'm disinclined to admit a complete epistemic/causal dichotomy, as I think it is well-established that we rely on all kinds of causal relations to perform epistemic functions (e.g., we appeal to the causal relations in our scientific instruments to show that they are reliable). I also think that suggestion has clear epistemic features independent of causal events. Suggestion is an ability in scientists that is trained, it is something we evaluate agents as being better and worse at. Making good suggestions is a kind of agential epistemic virtue. It's also not entirely the case that suggestion is a *mere* historical fact. Whether E is *taken* to suggest H is such a historical fact, but so is whether it is *taken* to *support* H. On the other hand, we can, looking back at the historical case, verify or evaluate whether E suggests H (this is probably clear to anyone who has had that "detective story" experience when reading some well-written history of science, coming up with the hypothesis "before" the scientist), just as we can do with support. We can also identify alternative hypotheses the evidence suggests that went unrecognized.

Finally, one might admit the epistemic relevance of the various functions I've identified, but nevertheless want a distinction between those functions and the narrower traditional understanding of evidence. Call the former "data" and the latter "evidence." It is tempting to take this sympathetic suggestion as a friendly amendment. As implied at the end of §5, various putative facts or items of evidence are used to make positive progress in inquiry which are eventually discarded, while at the end of an inquiry, we have reached a settled body of evidence which supports the final hypothesis. So there is a relevant distinction between what we might call "working facts" and "settled facts," which we might mark by making a distinction between data and evidence.

That the distinction exists is important to note independent of the purely terminological question of what words to use to mark the distinction. However, I would resist this particular terminological suggestion insofar as it suggests, falsely as I've argued, that one can have a theory of evidence worth the name without considering *both* working and settled facts and the connections between them. That is, one cannot understand the final body of evidence alone without understanding the functions it had to serve in inquiry in order to become the final body of evidence. Thus, we are less apt to fall into confusion if we talk about how a body of evidence is transformed in inquiry rather than distinguishing between two types of thing, "data" and "evidence." When we need to mark that distinction, I suggest we simply temporally or functionally index the body of evidence, e.g., the body of evidence in inquiry ongoing, the final body of evidence, the evidence guiding problem-formulation, etc.

My purpose in this part of the paper has been to motivate the use of more complex, temporally dynamic, functionalist models of evidence, to provide one such model, and to lend it some plausibility by showing how it would treat an illuminating example of scientific inquiry, indicating what I perceive to be its benefits, and responding to some objections against such a view. The model is beneficial in that it provides a realistic and plausible account of scientific practice that avoids some of the problems of the gross oversimplications in traditional models of evidence; it is nevertheless sufficiently general to provide some understanding of science and explanation of which strategies work well, and, as I will argue subsequently, it provides the strongest response to problems associated with the empirical basis of science. If I have oversimplified the nature of scientific evidence in turn, all for the better, since an even more complex account of the development of inquiry in time and the variety of evidential functions will serve my purposes just as well, if not better, so long as it remains manageable. With such an account in hand, the next step is to show how it can better cope with a variety of problems of evidence.

Part II
From the Experimenter's Regress to Evidence-Based Policy

7 Problems of the Empirical Basis of Science

7.1 Theory-Ladenness

The problem of the theory-ladenness of observation[27] is actually several problems with the following general form: are our observations (or observational evidence) infected with theoretical assumptions in such a way that undermines their ability to stand *as* evidence? Does the infection compromise the evidence's neutrality, objectivity, or reliability?

The first problem is the *linguistic* problem of theory-ladenness. This concern is most clearly connected with logical empiricism (or the so-called "orthdox" view), in which it was thought that a theory-neutral observation language was needed in order to adjudicate between competing theories. Philosophers like Popper, Kuhn, and Feyerabend were successful in dismantling this linguistic program in part by posing this version of the problem of theory-ladenness (though they differed greatly in their positive theories of science). The problem they present is that the language of observation, the concepts in which our claims about empirical evidence are couched, are themselves ineluctably laden with the assumptions of some theory. Concepts and terms as complex as "electron" or as simple as "length" cannot be independent of the context of some theory. Thus, if what is required for observations to stand as evidence of one theory over another is that they be couched in a neutral observation-language, we are presented with a serious problem.

Fortunately, almost no one today would agree with the empiricists that such a language is necessary. In the terms developed earlier, many philosophers have accepted some form of functionalism about evidence as against essentialism. Anything, regardless of the source of its concepts, can stand as evidence so long as, e.g., it can support or falsify hypotheses. The problem remains, however, if the structure of our observation reports builds in specific claims of our theory, in a way that would lead us to reject the substance of the report if we reject the theory. Whether this is ever really a problem is a matter of some debate.

Second, there is a *psychological* problem of theory-ladennes, whereby what one *sees* is conditioned psychologically by what one believes or accepts, or, by the skills and practices that one was acquired. In Hanson's famous example, one scientist looking at a figure immediately sees an X-ray tube; this is a product of what he's learned (his beliefs and skills). A baby would see no

[27] Peter Godfrey-Smith's way of describing the problem in his textbook *Theory and Reality* (2003, pp. 155–162) is felicitous, and while I disagree with his evaluation of the problem, I follow his way of setting it up in some respects.

such thing, nor would a member of a culture that was naive about such scientific devices. All seeing is seeing-as; there is no theory-neutral observation. Again, some forms of the problem are dropped when we adopt functionalism. However, if the structure of our observations is changed in a substantive way by our theory, hypothesis, or paradigm, the problem may recur.

Lastly, there are *procedural* or *instrumental* problems of theory-ladenness. Some of these are not a serious problem for most philosophers, as when a theory suggests where to look to find evidence that supports it. This might trouble one for a moment if you thought that a theory was a strict construction out of prior observations, but such an account of science has been so long out of fashion as to make it hardly worth mentioning except for its historical import. Only somewhat more serious is that theoretical assumptions generally guide experimentalists in discarding "erroneous" data, filtering out "noise," and so on. This has presented real problems in the history of science (e.g., the slowness to correct errors in measurements of fundamental quantities (like the charge of the electron) identified by Feynman (1986)), but such things tend to be corrected in the long run by ordinary scientific activity.

The most serious version of this problem is when the workings of an instrument presuppose the truth of some theory. For example, some authors (e.g., **?**, p. 82-3) have argued that Galileo's research with the telescope is not entirely rationally compelling because his theory of optics was very poor. On both Aristotelian and neutral grounds, it was reasonable to suppose that the telescope may be distortive rather than providing reliable evidence. Likewise, if some more recent apparatus presupposes the truth of some theory of the electron in order to regard it as producing useful data, such that if the theory is false, there would be no reason to regard the data as anything but noise, we have a significant problem of theory-ladenness. This is especially poignant when such a device is used to test the theory that its construction assumes; in such cases, we may well worry that the theory-test is just a complex way of begging the question.

Most traditional presentations of the problem of theory-ladenness assume essentialism and absolutism about evidence. As in the linguistic case above, where we drop the need for a theory-neutral language so long as observation reports are still able to confirm or falsify hypotheses depending on the contribution of the world, adopting functionalism significantly ameliorates the problem. Likewise, if evidence is tied to a context of assessment, then the context in which that evidence is used to assess some theory may have no problems, despite some other context in which the theory that informs that same bit of evidence being considered questionable. If, relative to some hypothesis the evidence is significantly more certain, it may well be able to stand as evidence.

Even if we assume the truth of a functionalist theory like Bayesianism, we may nonetheless have a form of theory-ladenness problem, to wit: the credence assigned to some evidence may depend on background assumptions that include the truth of some theory – in the worst case, the theory to be tested. If we look again to the instrumental problem of theory-ladenness, my evaluation of the standing of the observational evidence may depend crucially on my understanding of how the apparatus that produces that evidence works. If the theory is false, or if it is part of what I'm testing, that ought to undermine the credence I place in the evidence. This moves us towards the more subtle ways that mono-functionalism and non-dynamicism contribute to the problem.

According to the traditional model, the one and only function of evidence is to stand as the support for less certain theories or hypotheses. If a theory or hypothesis informs or creates that evidence, then the arrow travels in an illicit direction from theory to evidence, undermining the ability of the evidence to stand *as* evidence. The traditional model has little interpretive room to treat this as anything but a vicious circle. In some cases, merely lending terms or concepts to the observations may be innocuous, as perhaps the result may still falsify the theory. But if the theory also lends structure or credence to the evidence, the circle becomes problematic. Likewise, if we do not index judgments about evidential support to stages of inquiry, then we are forced to ask questions like, How does the body of accumulated evidence support the body of scientific theory? Then the problem of theory-ladenness becomes exacerbated by Quinean worries about holism of testing in familiar ways. It seems clear than in this case, skepticism may be the only result.

Many responses have been made to the problems of theory-ladenness, both empirical and philosophical, so many as to make it impossible to rehearse even a small sample here, and I do not mean to suggest that DEF is the only promising solution to the problem. Many such responses point out that there are reasonable ways to judge the power of evidence independent of the theory that "infects" it. (E.g., The reliability of Galileo's telescope was more a matter of trial and error than relying on bad optics, it tended to be accurate where independent checking was possible, and it provided seemingly coherent results rather than mere noise.) Such responses tend to go "off script" from the traditional model, which is no problem – until, moving on to some other issue, the problematic features of the model reassert themselves. The strongest response to the problem is thus to present an alternative model in which the facts about theory-ladenness present no problem.

Below, I will try to show how DEF treats the problem. I think this is a particularly strong response because (i) it provides a principled alternative to the approach that causes the problem, and (ii) it incorporates in unified fashion many of the reasonable but more *ad hoc* responses. But first, I want to follow a link between the problem of theory-ladennes and a related but apparently distinct problem, that of the experimenter's regress.

7.2 The Experimenter's Regress

Philosophers like Sylvia Culp (1995) worry about and attempt to solve the problem of the "experimenter's regress" raised by Collins (1975, 1992). Rather than a concern about how theoretical frameworks infect data, the experimenter's regress is a worry about how our expectations about results and our assumptions about certain techniques lead to circularity. According to Collins, what counts as good data is what results from a good experimental technique, but the test of an experimental technique is whether it produces the expected data. Collins, looking at the case of gravity wave detection experiments (1992, pp. 79ff), argues that,

> What the correct outcome is depends upon whether there are gravity waves hitting the Earth in detectable fluxes. To find this out we must build a good gravity wave detector and have a look. But we won't know if we have built a good detector until we have tried it and obtained the correct outcome! But we don't know what the correct outcome is until... and so on *ad infinitum*. (p. 84)

We have here a tight couple between the technique we use to gather data, the validity of the data itself, and our expectations about what data we should find. The "experimenter's regress" has two forms for Collins: a practical and a philosophical form. In the practical form, it presents a problem for scientists who must find a way to break the circle in order to resolve a dispute. In some cases, like the case of the TEA-laser that Collins discusses earlier in the book, the circle is broken by some practical result, e.g., the laser actually performs the required function. In the gravity-wave case, no easy external criterion (such as laser output of a certain kind) is available. Collins shows how variously interacting arguments about calibration, results, instrument sensitivity, assumptions about the data, the existence of the waves, etc. eventually led to the kind of "control on interpretation" that breaks the circle.

But from a philosophical point of view, this doesn't settle the problem. It is not on the basis of some conclusive evidence that the circle gets broken, but rather,

> the definition of what counts as a good gravity wave detector, and the resolution of the question of whether gravity waves exist, are congruent social processes. (p. 89)

And further,

> I am arguing here that just as the process of deciding whether gravity waves had been detected was coextensive with deciding which set of results was to be believed, so the detailed *nature* of gravity waves was settled at the same time. Different decisions

> about the quality of the experiments would have gone hand-in-hand with different decisions about the nature of gravity waves. (p. 100)

Since these decisions are made as a package, it is the *contingent*, *social* process of negotiation and decision-making that "break" the regress. The solution to the problem is thus a "sociological" rather than a philosophical solution (pp. 145ff), since experiments and evidence cannot do so. It is not that the way the dispute was settled is not reasonable or rational, just that a very different settlement would have been equally reasonable or rational. This leads to a form of *relativism* (p. 1) which holds that science studies should "treat descriptive language as though it were about imaginary objects"(p. 16) since it depends on contingent decisions, which different "networks of science and of society"(p. 130) would have made differently.

An important part of the problem of the experimenter's regress is the issue of *calibration*.[28] Early attempts to detect or measure some previously unobserved or unquantified phenomenon are faced with a problem of how to calibrate, lacking any other techniques to check against. We have only theoretical expectations about what the phenomenon should be like to guide us.[29] Later attempts are faced with the problem that their calibration depends on previous measurements which themselves were not calibrated in a standard way. In both cases there is a troublesome regress; in the earlier cases, we accept the measurement because it gave us the kind of results we expected – but then, it is hardly independent evidence for those expectations. In the later cases, we accept a measurement because it accords with our previous techniques in overlapping domains – but then, it is neither independent evidence for the reliability of our prior techniques, nor ultimately for our theoretical predictions.

7.3 The Justification of Evidence

Let us step back and think about the ideas about evidence in play. Something like the following picture, suggested by Culp (1995, pp. 439–40), is surely right: we set up an observational/experimental apparatus and run it. At one level, it merely produces brute happenings of a certain sort. We must then interpret those happenings, take them up as a certain item of fact, and, metaphorically speaking, teach them to speak the language of the theory, in order to see how they bear on the theory. (Of course, these "interpretations," according to certain defenders of theory-ladenness, take place at the level of

[28] See Franklin (2007, §I.B.1)

[29] Hasok Chang's work on temperature (2004) show in that case that there is ultimately a set of unchallenged expectations that inform what counts as an acceptable measurement technique. Basic assumptions like linearity, single-valuedness, etc. are inescapable (pp. 90–2).

seeing itself, not afterwards.) This interpretation may never be independent of theory, neither the theory of how the apparatus works nor the theory in question.

Further, thanks to the experimenter's regress, it is not only because we have a background *theory* informing our observation that data is infected, but more basic expectations about what data should look like and views about which techniques are reliable lead to a problematic circularity between data and technique. Sylvia Culp thus prefers to call Collins' "experimenter's regress" the "data-technique circle"(Culp, 1995, 438–9).

All of this presents a problem: we are left wondering how interpretations of experiments that *themselves* presuppose controversial theories, including parts of the theory in question, can serve as solid ground to support our theories; we are left wondering how claims about the reliability of a detector, which *themselves* presuppose controversial assumptions about what counts as "competent" data, including assumptions about the existence of the object in question, can serve as solid ground to support detection-claims.

The problem is one that is left unanswered in traditional models of evidence: how to we justify the evidence itself? In classical, foundationalist empiricism, the question is nonsensical or trivial: evidence is that which justifies but itself requires no justification.[30] Most philosophers have seen that classical empiricism is untenable. Nevertheless, they retain an imperfect allegiance to the traditional model which gives them few resources to answer the question of how evidence is then itself justified. It should be no surprise, then, that classical skeptical problems of regress or circularity loom very near.

7.4 The Robustness Solution and the Incommensurability of Evidence

One solution to the problems of the justification of evidence – theory-ladenness, the experimenter's regress, the data-technique circle – is to appeal to the *robustness* of evidence.[31] This solution is forcefully pursued by Culp (1995) and others. We need not have full independence of evidence from our expectations. Rather, what we need is evidence from a *variety* of different kinds of sources that are independent from each other and that still support the same conclusion. Evidence from a single source that seems to support the conclusion but only does so because it is calibrated to produce only supporting answers

[30]Except, perhaps, a general epistemological one. But the sort of justification in question is the justification of *particular* bits of evidence.

[31]Another interesting proposal is Chang (2004)'s *progressive coherentism.* Chang essentially accepts the circularity, arguing that the self-correcting nature of science prevents it from being vicious. This solution may be adequate in cases like temperature, where some basic independent grip on the phenomena in question (like perceptual awareness of hot and cold), but it seems more problematic in cases like the ones Collins worries about, where multiple progressive circles are allegedly possible or actual.

would be problematically circular. A variety of different types of evidence, developed independently from each other at different times and places, which all seem to support the conclusion but in fact are just the product of our expectations, so the argument goes, would be a miracle – or at least an "improbable coincidence"(Culp, 1995, 448). A common cause "other than shared theoretical presuppositions"(*ibid.*) – the truth of the conclusion – is the better explanation.

The strategy is an appealing one. Suppose you want to build a bridge to carry a train across a ravine. All the individual wooden boards at your disposal, taken separately, are inadequate to carry the weight of the train. One could either give up on the possibility of using wood to support the train, or one could try to figure out if a large enough collection of boards, arranged in a very particular way, might do the job. Culp (1995) fully admits that no particular bit of evidence can be theory-free or assumption-free, that it doesn't even make sense to talk of uninterpreted, bare "happenings" as evidence (439). Nonetheless, since she is committed to the metaphor of support, she attempts to find an arrangement of evidence that can serve as a fixed-enough support. A set of evidence can be a foundation for theoretical knowledge if it is *robust* – if it comes from a variety of sources that are theoretically independent of each other.

This argument unfortunately fails to meet the challenge posed by the experimenter's regress. At least three difficulties arise, one empirical and two epistemological.[32] The first is the difficulty of finding *really* independent sources of evidence. The history of the development of experimental techniques is replete with a variety of cross-calibration techniques. Hasok Chang's (2004) discussion of the development of the modern thermometer shows the complex interdependencies of various new techniques for measuring temperature (see especially Chapter 3). Early errors propagate into later techniques and take a long time to disappear entirely, as in the case of measurements of the charge of the electron (Feynman, 1986), because of the preponderance of cross-calibration. True independence may be difficult to determine (Stegenga, 2009, 652-4).

The second problem, which springs from the first, is that robustness doesn't really solve the problem of calibration. For any particular measurement technique, there are two cases: either it is calibrated according to existing techniques, or it isn't. In the former case, the possibility of independent techniques of measurement is seriously endangered. Furthermore, the question of how those pre-existing techniques were themselves calibrated must be examined. In the latter case, it would appear that all we have to go on to judge the results provided by the technique is the very expectations we hope to support. A variety of different types of evidence, all calibrated by reference to the same set of expectations also lack the independence required by the argument.

[32]My argument here is strongly influenced by Stegenga (2009).

It may be that the original types of measurement, though originally calibrated in a suspect way, are calibrated with respect to different, independent sets of expectations. While problematic in those original circumstances, in a *present* case, they may be sufficiently independent *from one another* to provide robust, adequate evidence in the case at hand. Even supposing that this case passes the empirical test of independence discussed above, a larger epistemological question about whether we ought to rely on the evidence remains. Perhaps we ought to regard it as a miracle that a variety of such evidence purportedly supports a single conclusion, but why should we think for one moment that the truth of that conclusion explains the apparent miracle, given the story of evidence now on offer?[33] A variety of methods, calibrated under highly suspicious circumstances, apparently providing no trustworthy support in the case of their original development, now all happen to agree on one conclusion. Do we have any reason to believe that this coincidence has anything to do with the truth of the conclusion? Not without some prior reason to think that the methods, taken individually, track the truth in even a modestly reliable fashion, i.e., that the methods track *some* signal, and don't just produce noise. But it is *precisely* the lack of such a reason in the case of individual techniques that leads to the demand for robustness in the first place. The coincidence found in patterns of pure noise is just that – a coincidence. In terms of our metaphor above, we can build a strong bridge out of comparatively weak individual boards, but not out of wet noodles. Robustness cannot distinguish the cases.

The final problem is the nail in the coffin for the prospects of solving this problem through the appeal to robustness. In order to have truly independent sources of evidence, it is crucial that the measurement techniques not be calibrated to one another, lest the bias in one creep in to the other. The sources must be of different types, and they *must* be *incommensurable*, in the sense of not having any common standard of comparison. This is so because the existence of such a standard implies mutual calibration. In order to ensure that the sources of evidence are *really* independent, we must avoid theoretical connections, reliance on shared assumptions about the physical processes at work, and cross-calibration. The existence of standards of comparison means that we either have cross-calibrated the techniques, or we have relied on some shared theory or assumptions about how the processes work. Either invalidates the requirements on robustness implied by Culp's argument.

If they are incommensurable in this way, however, we're left with a major worry: if we have no standard of comparison between the types of evidence, how can we say determinately that they support the *same* conclusion? If the interpretive framework at hand is the theory in question, of course, then it is easy to see how different bits of evidence support the same conclusion. But

[33] General reasons to be skeptical of no-miracles arguments (and their anti-realist counterparts, the pessimistic induction) are provided by Magnus and Callender (2004).

then the evidence isn't really independent in the way that Culp demands. Suppose, then, that the evidence, that is, "raw data" plus interpretation, are all independent from one another. How do you determine the relevance of each to your hypothesis?

This question may be practically answerable in a relatively loose and informal way, when all of the evidence seems to tell in favor of, or is at least consistent with, the hypothesis. But what if the evidence isn't so concordant? Indeed, the conditions which were meant to solve our problem and break us out of the data-technique circle have created an even more difficult problem: the problem of incommensurable but *discordant* evidence.[34] This problem will be addressed in §8. For now, I will take for granted that the seriousness of this problem is sufficient to justify the attempt an an alternative to robustness *as a solution for the problem of justifying evidence*.[35]

7.5 Evidence, Hypothesis, and Functional Fitness

Considering these issues from the point of view of the DEF-model of evidence, the accounts of evidence used in discussing these problems of evidence seem rather impoverished. For one, they mention only one direction on the two-way street of the coordination of evidence and hypothesis. Contra Culp's supposition, we don't only teach evidence to "speak the language" of theory. We also teach the theory to speak the language of observation; that is, we must develop our hypotheses so that they have operational consequences, that they may *direct* activities of observation. This too is an "interpretation," if you like, of the theory, but it is very different from the process of interpretation that Culp discusses. Collins' and Culp's shared way of setting up the problem of the justification of evidence presupposes that hypotheses are inert, and experiment must be constructed or interpreted in a way that meets it. But hypothesis and experiment must meet in the middle.

Further, they construe the function of evidence extremely narrowly. Evidence is taken to be exhausted by its function of supporting a hypothesis. According to DEF, this is just one functional role amongst many. It is undeniable that in some sense, theories, hypotheses, ideas "produce" their own evidence. But this is only a problem if evidence serves only to justify, and theory or hypothesis is justified only by that body of evidence it produces. To the contrary, producing (not predicting) some events is the *point* of a theory or hypothesis (or one of them); it is the adequacy of the consequences produced to solving the problem, along with its usefulness in attacking *new* problems

[34] The same conclusion is reached by Stegenga (2009). His argument is somewhat different, and I believe I draw a stronger conclusion about the demands of robustness requiring incommensurability of evidence.

[35] Outside of the attempt to use robustness to solve the problem in a way that saves the traditional model, which leads to very strong requirement on the alternative forms of evidence, robustness is a very useful concept indeed.

and supplementing *new* inquiries, that are the ultimate test of the theory. A theory which *failed* to produce its own evidence, i.e., failed to produce any new phenomena, would be inert, useless, and unjustifiable. It would be impotent to solve any problems.[36]

Evidence has a variety of functional roles within an inquiry, the main goal of which is the resolution of the perplexity which spurred the inquiry. In general, then, the experimenter's regress will not present any difficulty, since all that matters is that the evidence fulfill its role *well enough* for the purposes of solving whatever problem presents itself. In other words, evidence and hypothesis (theory) alike are ultimately adjudged according to their *functional fitness* in problem-solving inquiry.[37]

Genuine problems of inquiry set the conditions of their own solution. They do not go away because some external standards of "objectivity" or "justification" are satisfied. Only a transformation of the situation to remove the original discoordination or difficulty (perplexity) will suffice. Since experiment is *not* merely a procedure for producing neutral evidence, but *rather* a way of *making* and *doing* that puts the hypothesis *into practice*, there is a test of the experimental evidence, together with the hypothesis, that is independent of expectations *per se*. Expectation cannot prevent a bridge from falling down, nor can it cure disease, nor can it even reconcile the incompatibility between quantum mechanics and relativity theory. The germ of this solution exists in Collins' own discussion, i.e., in the case of the TEA-laser. What Collins ironically misses is that the question of the existence of gravity waves is not itself a context-free, abstract question, but rather part of a social process of dealing with a problematic practice (a perplexity), and the concrete factors of that situation provide the conditions for adequate solution, just as the narrowly *practical* function of the laser provide the conditions for adequate solution in that case.

8 Problems of Combining Evidence

In recent work, Jacob Stegenga (2009) has discussed the problem, raised by Franklin (2002), of discordant evidence, that is, the problem of how to address diverse, multimodal evidence which appears to pull towards *different* conclusions. For example, Stegenga discusses the case of the transmission of influenza. Clinical evidence such as patterns of transmission suggest that the flu is transmitted only by contact. On the other hand, mathematical models and some case studies suggest that it is quite likely that the influenza virus is spread through the air. Given the (as I've argued, necessary – for the ro-

[36] Though it is important to keep in mind that the range of problems and the diversity of ways that phenomena are produced exceeds what is commonly called "practical" in the narrow sense.

[37] Cf. Dewey (1938, *Logic*, LW 12:114).

bustness theorists) lack of any meta-standard for balancing diverse evidence, difficult decisions must be made about which set of evidence is more relevant in this case. The problem of discordance not only raises doubts about the value of robustness, but raises a clear problem for scientific methodology itself: if evidence of different types conflict, what are we to do when making decisions where evidence is required?

When evidence is of one type only, fully commensurable, problems of discordance do not occur. There may be disagreement between results, but these can be chalked up to error, noise, or a problem with the technique. It is a basic assumption of a measurement technique that it provides consistent results within its margin of error (Chang, 2004, 90–2).[38] Further, when different techniques are commensurable, as in the measurement of temperature with a wide variety of thermometers (Chang, 2004, Ch. 2 & 3), it is common practice to calibrate the techniques so that they give consistent results when their areas of functioning overlap. When techniques are multi-modal and incommensurable in the way that *robustness* requires, however, the problem of discordance arises. Franklin (2002) suggests that robustness can *solve* this problem, but, as Stegenga argues (and as I do above), it is the requirement of robustness that evidence be multi-modal and strongly independent that causes the problem. In cases where evidence is not incommensurable in the sense that robustness requires, no particularly difficult problem of discordance need arise. In cases where evidence aims at robustness, discordance will often arise, and cannot be erased by gathering further evidence.

Appealing to robustness alone, the best one can do is increase the amount of evidence pointing in one direction. This fails as an adequate solution to the problem of discordance, however, as it fails to address what Cartwright (2009) and Stegenga (2009) term "the problem of relevance." When multi-modal, incommensurable evidence disagrees, it matters not only what the *quantity* or even the *quality* of the evidence is. It also matters which evidence is more *relevant* to the problem at hand. In the epidemiological case mentioned above, much of the controversy depends on one group believing that the clinical evidence is more relevant, while others think that the models and case studies are. This goes beyond mere precision and validity. The question is, given a hypothesis, which evidence bears more directly on its truth or falsity.

If, as the adherents of robustness must admit, the hypothesis and all the different types of evidence must come from independent conceptual backgrounds, and thus to some degree "speak different languages," then the problem of relevance is of upmost importance. We must be able to determine how some piece of evidence bears on some hypothesis where there is no simple way to plug

[38]Of course, more sophisticated measurement techniques than thermometers may produce evidence that appears less consistent, and statistical analyses must be applied to make sense of the results. But then, I would say that what functions as "evidence" in this process are not the individual data-points that are fed into the analysis, but the original process of analysis itself. My thanks to Jacob Stegenga for reminding me of this complication.

them in to a probabilistic formula nor a deductive syllogism. We must figure out how to reckon three or more independent factors: the hypothesis and the different forms of evidence.

But as soon as we state the problem this way, it seems utterly insoluble. If there is no common ground between putative pieces of evidence, or between evidence and hypothesis, how can they be reconciled? Without standards for mutual comparison of the type that vitiate independence, what way is there to settle the differences?

One possibility is to find a formal meta-standard for comparing evidence and determining its relevance that is independent of and blind to the background assumptions in question. Such standards are in place for so-called evidence-based policy which only look at experimental design (RCT, case-control study, etc.), but these fail to really capture relevance (see §9 below). In general, such standards will fail because the problem of relevance depends on the *content* of the evidence and hypothesis, not just the formal aspects. Furthermore, we should generally be suspicious of such formulae; the attempt to find a simple algorithm or recipe for reconciling various types of evidence amounts to an attempt to solve a difficult task faced by all research in one fell swoop. In all likelihood, this is simply a problem that must be solved in the course of each inquiry, on its own terms, and cannot be eliminated by philosophical sophistication.[39] We might be able to discern some interesting generalities about the strategies by which scientists do so, but we should not expect a short-cut solution.

Perhaps, while no formal methods of reconciliation are available, good scientists will nevertheless be able to see how to determine the relevance of the evidence to a hypothesis. Science is a creative, skillful activity, and while no explicit rules can be articulated, the tacit knowledge available to practitioners allows them to make good judgments about relevance. While this must to some degree be correct, it is an inadequate answer to the problems of relevance and discordance. First, it is difficult to normatively assess tacit knowledge and skilled judgment. There is a difference between what judgments scientists are *justified* in making and simply what a scientist or group of scientists *in fact* does, but it isn't clear how to distinguish the two if scientific judgment is so inarticulable. Second, this doesn't address the way that disagreements about relevance and how to resolve discordant evidence are pervasive in scientific controversy. If skillful judgment can resolve the problem, then why is there so much disagreement on just this matter? Finally, this answer presupposes an illegitimate individualism in its understanding of the scientific process. Ultimately, it is not individual scientists who have the last word on scientific debates. Rather, science is a social phenomenon, and these matters must be decided on a larger scale than individual judgment. Skill and tacit knowl-

[39]Such attempts at short-cut solutions are a vicious temptation in philosophy, especially epistemology. More on this below.

edge surely play a role in how science gets done, but settling disputes over discordant evidence must take place at a more explicit level.

What we must do is reject "robustness" altogether (in the very specific sense that defenders like Culp are forced to accept). The call for evidence that is independent from the hypothesis in question and a set of evidence each independent from the rest is an impossible requirement. Without some shared background and structures of commensuration, without the ability to coordinate hypothesis and evidence, there is no way to push inquiry forward.

Which is not to say that there are never difficulties of determining the relevance of some data, or that there are never problems of inconsistence or incongruity between evidence. Discordance *can* be a real problem, not for epistemology but for scientific inquiry itself. What is problematic in the way that some philosophers approach evidence is that they hope to solve this problem once-and-for-all with some formal method or meta-standard that obviates the need for further research. Franklin's instincts are right when he suggests that the problem of discordance can be solved by gathering further evidence, but this answer fails if we understand it either in terms of robustness laid out above or in terms of the traditional models of evidence.[40] Looking at the problem of discordance from the point of view of the DEF-model, resolving discordant evidence is just an ordinary part of resolving the inquiry in question.

At the beginning of an inquiry, we expect discordant evidence. If the evidence was at first blush all in agreement, there would be no problem for inquiry to resolve. The situation would already be settled (at least in the relevant respects), and so we could simply apply our theory and move on. Discordant evidence is part of what sets the problem for inquiry in the first place, because it suggests conflicting possibilities. When Snow first began to study cholera,[41] there was evidence that pointed towards it being an ordinary, communicable disease, but there was also evidence that many people exposed to the disease, such as doctors, rarely caught it, and others never exposed to cholera patients nevertheless caught the disease. How then is the disease transmitted and how can it be contained? Discordance will also naturally arise in the mediate phases of inquiry, and it is a driving force for the improved articulation of both the hypothesis and the data. Snow was able to explain with his direct-ingestion hypothesis why doctors rarely contract the disease. But then he had to explain the occasional epidemics in which the rich and poor alike became infected. This drove further articulation of the hypothesis (transmission via water supply) and suggested new observations and experiments (track and intervene in the distribution of water).

[40]Furthermore, it seems to me that Franklin's own views, while more rationalist than I am inclined to, diverges in many respects from the traditional account. Franklin is not so much beholden to the traditional philosophical frameworks as he is simply trying to explain the successful strategies of scientists (especially experimental physicists) in the face of concerns raised by skeptical philosophers and sociologists about the authority of science.

[41]See Goldstein and Goldstein (1978, 25–62).

Discordant evidence is part of the problem of inquiry, because it sets the problem and is part of the mediate phases of inquiry in which the attempt to coordinate evidence and hypothesis is not yet complete. It is not a *further* problem for epistemology, if that means that we should be looking for a way of resolving it that goes beyond the way it is done in the ordinary course of inquiry. Take the example of influenza again. Controversy continues about whether it is airborne or transmitted only by contact, with each side marshalling evidence in its favor. If we ignore the temporal complexity of inquiry, then we can see this as a serious problem for epistemology: given this conflicting evidence, what should we believe, what should we do? If we attend to the process of inquiry, however, we see that this is simply an intermediate phase of the investigation. Philosophers cannot settle it by fiat; scientists must settle it. And they must do so by proposing and refining hypotheses that explain[42] the discord, finding reasons to reject apparently relevant evidence, gathering further evidence and constructing new experiments that bring the controversy to a close.

One might respond that while scientists might have all the time in the world to settle the theoretical question of the nature of influenza, decisive action must be taken *now* to control and prevent this sometimes life-threatening disease. So it must, but action and policy are not separate from the process of inquiry, as I will discuss in the next section.

John Dewey indicted much of traditional philosophy for attempting a short-cut around inquiry when only inquiry would do, a quest for certainty which is misconceived at best and positively damaging at worst. Philosophers have been at their best when they observe and distill the lessons of inquiry so that they may be made available in other inquiries, though even this often happens *despite* their intentions to seek certainty.[43] In the current discussions of evidence, the temptations to resist are (1) a short-cut around the difficult task of inquiry and (2) to declare the problem insoluble.

9 Problems of the Application of Science

Evidence-based policy (EBP) is a fast-growing movement in public policy, especially in the areas of medicine and education. Already, plenty of government funding, hospital policies, and educational mandates hang on the existence of

[42]Heather Douglas has recently argued for and *explanation* solution to problems about weighing and integrating evidence. It is not some special formal framework, but the ordinary canons of good scientific explanation that do the job. See Douglas (2010, §5) and the talks from December 2009 and April 2011 on `http://utk.academia.edu/HeatherDouglas/Talks`

[43] See the final chapter of the *Logic*, "The Logic of Inquiry and Philosophies of Knowledge," where Dewey discusses the ways in which traditional philosophers of different schools have, by partial attention to only some features of inquiry, gotten certain aspects right and others wrong. (Dewey, 1938, 506–527)

certain kinds of evidential standards. In practice, this means that policies are funded or approved on the basis of whether there exists evidence for the policy that ranks highly on one of the prominent evidence-ranking schemes, such as SIGNS (the Scottish Intercollegiate Guidelines Network) and the "What Works Clearinghouse" of the US Department of Education. These schemes inevitably put randomized controlled trials (RCTs) at the top of the list, and things like case studies, ethnographic studies, and expert opinion either aren't mentioned or are ranked very low (Cartwright and Efstathiou, 2008).

Philosophers have raised a variety of objections. For example, John Worrall has argued that EBP has overestimated the value of RCTs and underestimated the value of expertise, because it pursues an unrealistic strategy of attempting to eliminate alternative explanations without making any judgments of what counts as a *plausible alternative* (Worrall, 2002). Nancy Cartwright has argued that EBP lacks justification because we lack any "reasonable and practicable" theory of evidence that could do the work EBP requires. The standards in place are too restrictive, make plain wrong claims about strength of evidence, and provide no useful information about combining evidence. (Cartwright, 2007). The standard rankings also evaluate soundness about evidence without providing information about whether the evidence is sufficiently *relevant* to the policy in question (Cartwright, 2009). These are some examples of prominent criticism, and once one gets into the details for particular areas, e.g., of medicine or education, the criticisms multiply.

What can we say about EBP from the point of view of the DEF-model? As everyone in the discussion is quick to point out, *of course* basing policy on evidence is a good thing (though one often wishes this view were more consistently held amongst politicians). Nevertheless, the particular way that evidential standards have been drawn up suffers from the same worry that many of the discussions of evidence do, namely, it attempts to short-cut the need for research with an easy answer. Understanding this point of criticism in the policy case requires that we shift how we think of the nature of policy-making.

EBP is meant to do two things: it is meant to make consulting scientific evidence where questions of fact are relevant to policy-making *mandatory*, and it provides a standard of evidence *for policy* to assess the quality of evidence provided. In the real world, the sources of evidence are more diverse and complex than some monolithic Science, running the gamut from publically-funded research to corporate R&D, and the policy process itself is complex and adversarial. Since representatives of science *tout court* rarely convene to provide an univocal answer to policy queries, standards for assessing evidence in favor of competing proposals seem necessary.

The attraction of such a standard is obvious, as it makes the difficult process of weighing evidence susceptible to a fairly simple algorithm. But its sheer simplicity is cause for alarm. Consider an analogous case in scientific inquiry: when a controversy between two competing theories or explanation is in process, with each side marshaling evidence in its favor, the question cannot be finally answered, and inquiry brought to a close by the application of an algorithm. As we've seen, this is a question that requires *further inquiry*, and short-cut solutions won't do the job.

The shift that ought to be made in our understanding of policy in order to avoid the false certainty of a short-cut solution is to regard *policy itself* as an inquiry,[44] different in some ways from scientific inquiry, but inquiry nonetheless. Already the DEF-model requires that we regard science as something other than a mere accumulator of impartial information; rather, science is a problem-solving process that attempts to resolve a variety of perplexities, from the mundane and practical to the abstruse and distant from immediate application. The policy-process itself can be profitably understood as one of identifying and attempting to resolve social problems of a certain sort.

On the DEF-model, a general commitment to evidence-based policy is a no-brainer. Obviously policy, like any inquiry, *must* be based on evidence. But evidence doesn't come pre-packaged by other areas of inquiry. While the conclusions of other inquiries provide *prima facie* materials for further inquiries, the adaption of evidence into different contexts is never automatic, nor can pre-existing evidence be expected to be sufficient for resolving an inquiry. New evidence must be gathered on the problematic situation *at hand*, on the basis of the current perplexity. The relevance of old evidence must be determined by attempts to coordinate it with an understanding of the problem and proposed solution and on the ability to generate further evidence on that basis, congenial to solution. The validity of the evidence in a new context is always in question, susceptible to revision or rejection as the inquiry moves forward.[45] Ultimately, *policy itself* must be understood *not* as a final answer, but as itself an experiment which must be approached tentatively and taken as provide evidence about the adequacy of a proposed solution. We need to strive for *inquiry*-based policy rather than evidence-based policy.

References

Barnes, E. C. (2008). *The paradox of predictivism.* Cambridge Univ Press.

Bogen, J. (2010). Theory and observation in science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2010 edition.

Bogen, J. and Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, 97(3):303–352.

[44] See Kaufman-Osborn (1985); Caspary (2000).

[45] Cf. Bryan Norton's discussion of risk assessment in (Norton, 2005).

Bogen, J. and Woodward, J. (1992). Observations, theories and the evolution of the human spirit. *Philosophy of Science*, 59(4):590–611.

Bogen, J. and Woodward, J. (2005). Evading the IRS. *Idealization XII: correcting the model: idealization and abstraction in the sciences*, page 233.

Browning, D. (1994). The Limits of the Practical in Peirce's View of Philosophical Inquiry. In Moore, E. C. and Robin, R. S., editors, *From Time and Chance to Consciousness: Studies in the Metaphysics of Charles Peirce*, pages 15–29. Oxford: Berg Publishers,.

Cartwright, N. (2006). Well-Ordered Science: Evidence for Use. *Philosophy of Science*, 73:981–990.

Cartwright, N. (2007). Evidence-based policy: Where is our theory of evidence? Technical Report 07/07, Centre for Philosophy of Natural and Social Science, London School of Economics.

Cartwright, N. (2009). Evidence-Based Policy: What's To Be Done About Relevance. *Philosophical Studies*, 143(1):127–136.

Cartwright, N. and Efstathiou, S. (2008). Evidence-Based Policy and Its Ranking Schemes: So, Where's Ethnography? In *conference of the Association of Social Anthropologists The Pitch of Ethnography, LSE*.

Caspary, W. (2000). *Dewey on Democracy*. Cornell University Press.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press, USA.

Cohen, J. D. (2009). *The red and the real: an essay on color ontology*. Oxford University Press, USA.

Colaço, D. (2018). Rethinking the role of theory in exploratory experimentation. *Biology & Philosophy*, 33(5):38.

Collins, H. (1975). The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology*, 9(2):205.

Collins, H. ([1985] 1992). *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

Culp, S. (1994). Defending Robustness: The Bacterial Mesosome as a Test Case. In Hull, D., Forbes, M., and Burian, R. M., editors, *PSA 1994: Proceedings of the 1994 Biennial Meeting of the Philosophy of Science Association*, volume 1, pages 46–57, East Lansing, MI. Philosophy of Science Association.

Culp, S. (1995). Objectivity in experimental inquiry: Breaking data-technique circles. *Philosophy of Science*, 62(3):438–458.

Dewey, J. (1938). *Logic: The Theory of Inquiry*, volume 12 of *The Later Works of John Dewey*. Southern Illinois UP, 1991.

Douglas, H. (2010). Engagement for progress: applied philosophy of science in context. *Synthese*, 177:317–335.

Drengson, A. (2010). Four Philosophies of Technology. *Technology and Values: Essential Readings*, pages 26–37.

Drengson, A. R. (1982). Four philosophies of technology. *Philosophy Today*, pages 103–117.

Dweck, C., Chiu, C., and Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A word from two perspectives. *Psychological Inquiry*, 6(4):267–285.

Feynman, R. (1986). Cargo cult science. *Surely you're joking, Mr. Feynman*, pages 338–346.

Franklin, A. (1994). How to avoid the experimenters' regress. *Studies In History and Philosophy of Science Part A*, 25(3):463–491.

Franklin, A. (2002). *Selectivity and Discord: Two Problems of Experiment*. University of Pittsburgh Press, Pittsburgh.

Franklin, A. (2007). Experiment in Physics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.

Franklin, L. R. (2005). Exploratory experiments. *Philosophy of Science*, 72(5):888–899.

Giere, R. N. (2006). *Scientific Perspectivism*. University of Chicago Press, Chicago.

Glymour, B. (2000). Data and phenomena: A distinction reconsidered. *Erkenntnis*, 52:29–37. 10.1023/A:1005499609332.

Godfrey-Smith, P. (2003). *Theory and reality: an introduction to the philosophy of science*. University of Chicago Press.

Godin, B. and Gingras, Y. (2002). The experimenters' regress: from skepticism to argumentation. *Studies In History and Philosophy of Science Part A*, 33(1):133–148.

Goldstein, M. and Goldstein, I. (1978). *How we know: an exploration of the scientific process*. Westview Press.

Hacking, I. (1982). Experimentation and scientific realism. *Philosophical Topics*, 13(1).

Hey, S. P. (2015). Robust and discordant evidence: Methodological lessons from clinical research. *Philosophy of Science*, 82(1):55–75.

Hickman, L. A. (1998). Dewey's theory of inquiry. In Hickman, L. A., editor, *Reading Dewey: Interpretations for a Postmodern Generation*, pages 166–86. Indiana University Press.

Jackson, F. and Pettit, P. (1995). Moral functionalism and moral motivation. *The philosophical quarterly*, 45(178):20–40.

Jackson, F. and Pettit, P. (1996). Moral functionalis, supervenience and reductionism. *The Philosophical Quarterly*, 46(182):82–86.

Karaca, K. (2013). The strong and weak senses of theory-ladenness of experimentation: Theory-driven versus exploratory experiments in the history of high-energy particle physics. *Science in Context*, 26(1):93–136.

Kaufman-Osborn, T. V. (1985). Pragmatism, policy science, and the state,. *American Journal of Political Science*, 29(4):827–849.

Kelly, T. (2008). Evidence. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition.

Kuhn, T. S. (1962 [1996]). *The Structure of Scientific Revolutions*. University Of Chicago Press, 3rd edition.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. *Criticism and the Growth of Knowledge*, pages 91–195.

Laudan, L. (1977). *Progress and its problems: toward a theory of scientific growth*. University of California Press, Berkeley.

Laudan, L. (1984). *Science and values: the aims of science and their role in scientific debate*. University of California Press, Berkeley.

Levin, J. (2009). Functionalism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Winter 2009 edition.

Lynch, M. P. (2000). Alethic pluralism and the functionalist theory of truth. *Acta Analytica*, 24:195–214.

Lynch, M. P. (2001). A functionalist theory of truth. *The Nature of Truth*, pages 723–750.

Lynch, M. P. (2005). Alethic functionalism and our folk theory of truth. *Synthese*, 145(1):29–43.

Magnus, P. and Callender, C. (2004). Realist Ennui and the Base Rate Fallacy. *Philosophy of Science*, 71(3):320–338.

Martin, W. (2006). *Theories of Judgment: Psychology, Logic, Phenomenology*. Cambridge University Press.

Matheson, C. (2009). Historicist theories of rationality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2009 edition edition.

Norton, B. G. (2005). *Sustainability : a philosophy of adaptive ecosystem management*. University of Chicago Press, Chicago.

Oberheim, E. and Hoyningen-Huene, P. (1997). Incommensurability, Realism and Meta-Incommensurability. *Theoria*, 12(3):447–465.

Peirce, C. S. (1877). The Fixation of Belief. *Popular Science Monthly*, 12(1):1–15.

Stegenga, J. (2009). Robustness, Discordance, and Relevance. *Philosophy of Science*, 76:650–661.

Sternberg, R. (1985). Implicit theories of intelligence, creativity, and wisdom. *Journal of personality and social psychology*, 49(3):607.

Talbott, W. (2008). Bayesian epistemology. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition.

Woodward, J. (1989). Data and phenomena. *Synthese*, 79(3):393–472.

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science*, 69(S3):316–330.

Wright, C. D. (2005). On the functionalization of pluralist approaches to truth. *Synthese*, 145(1):1–28.