Decisions, Decisions: Inductive Risk and the Higgs Boson

Kent W. Staley

Saint Louis University[1]

Abstract:

Much of the discussion of the Argument from Inductive Risk (AIR) centers on scientific research that has relevance to policy-making. To emphasize that inductive risk pervades science, this chapter discusses the AIR in the context of High Energy Physics (HEP): specifically, the discovery of the Higgs boson, a scientific finding that is irrelevant to policy. The applicability of the AIR for the case of the Higgs boson is established through a pragmatic approach to scientific inquiry, emphasizing the centrality of practical decision problems to the production of scientific knowledge. This approach, drawing on debates among pragmatists over the interpretation of statistical inference, eschews the classification of value judgments into epistemic and non-epistemic.

Keywords: inductive risk, high energy physics, Higgs boson, pragmatism, statistics, discovery

1. Introduction

[1] A previous version of this paper was presented as part of a symposium on discovery in High Energy Physics at PSA 2014 in Chicago, Illinois. I am grateful to my fellow symposiasts Bob Cousins, Allan Franklin, and Deborah Mayo, as well as our session chair Vitaly Pronskikh, for fruitful discussion of these issues. I would also like to thank Ted Richards and Kevin Elliott for their helpful feedback on an earlier draft.

Applications of the Argument from Inductive Risk (AIR) typically highlight scientific research that bears obviously on policy matters. In this paper, I consider the AIR in the context of research lacking in clear policy implications: the search for the Higgs boson. The discovery of the Higgs boson in July 2012[2] unleashed no new technology that could enhance or harm human health; it revealed no environmental, economic, or political problem that would prompt contested proposals for redress. Rather, it contributed support to a proposed answer to a seemingly arcane, yet quite deep, problem of theoretical physics: The carriers of the weak force (the *W* and *Z* bosons) have mass. But it had been thought that particles carrying forces in theories possessing the formal property of *gauge invariance* (like the Standard Model) would have to be massless. Physicists introduced the Higgs field as a means to reconcile the experimentally established masses of the *W* and *Z* with the gauge invariance of the Standard Model (SM), and the Higgs boson is the particle predicted as an excitation of that field.

The details of the physics need not concern us (see Allen 2014 for a nice discussion aimed at a general audience); here I aim only to make clear just how remote from practical concerns the scientific question of the Higgs boson lies, making it plausible that the only relevant (or at least the dominant) values in this inquiry are those relating to belief, and not action (Myrvold 2012, 555).

In this paper I will argue that, nonetheless, issues of inductive risk are relevant in the discovery of the Higgs boson. At the same time, I introduce a non-standard perspective on what those issues are. My approach eschews the classification of value

[2] More precisely, this was the discovery of a new boson with properties consistent with those theoretically attributed to the Higgs boson in the Standard Model (SM) of particle physics, but whose identification as an SM Higgs awaited further data (Aad et al. 2012; Chatrchyan et al. 2012).

judgments into epistemic and non-epistemic. Rather, I return to the mid-twentieth-century roots of the literature on the AIR and extract from them a debate over how to apply a broadly pragmatic philosophical orientation to the interpretation of statistical inference. In so doing, I emphasize the central role of practical decisions in the production of theoretical knowledge.

The Higgs discovery involved two groups operating at the Large Hadron Collider (LHC) in Geneva, Switzerland: the CMS (Compact Muon Solenoid) collaboration and the ATLAS (A Toroidal LHC Apparatus) collaboration. Their announcements invoked a statistical standard for discovery claims in High Energy Physics (HEP) that I will henceforth call the *five sigma* (5σ) *standard*. Both groups claimed to have found evidence for a new particle with a statistical significance of five standard deviations (5σ) (Aad et al. 2012; Chatrchyan et al. 2012).

Relying on a statistical standard for the decision to accept a hypothesis is a crucial element in the arguments of both Richard Rudner and C. West Churchman (cited as early proponents of the AIR), a point recognized in critical responses such as Isaac Levi's. Focusing on Churchman and Levi, I cast their dispute in terms of the demands of a pragmatic account of scientific inference and the prospects for what Levi calls "epistemic autonomy." Drawing on the Higgs discovery as an illustration of how the *practical* permeates even the most policy-irrelevant inquiry, I will argue that, even if Isaac Levi's attempt to ensure epistemic autonomy succeeds in principle, its relevance for scientific practice is limited. A more thoroughgoing pragmatism is required in order to understand how the evaluation of experimental data contributes to the production of scientific knowledge.

Section 2 of this paper sketches the statistical methodologies of significance testing and Neyman-Pearson hypothesis testing, implicated both in the philosophical issues and the scientific case considered. Section 3 explicates the AIR and introduces Churchman's pragmatist project in inductive reasoning. Section 4 surveys Levi's defense of the epistemic autonomy of scientific reasoning. Section 5 discusses the use of the 5σ standard in experimental HEP and argues for a severe limitation of the scope of Levi's defense of epistemic autonomy. A conclusion summarizes and clarifies the thesis here defended.

2. *p*-values and error probabilities in HEP

By characterizing their evidence in terms of an estimate of the statistical significance of their findings, ATLAS and CMS incorporated the language and methodology of significance testing, a statistical methodology for testing hypotheses that utilizes probabilities understood as relative frequencies. Here I briefly summarize in a rough and informal way this widely used methodology.

A significance test is a device for answering a question. To attempt an answer, one formulates a substantive hypothesis that is a possibly correct and testable answer to that question, stated in scientifically meaningful terms, such as (in this case) the terminology of the Standard Model. This is the *null hypothesis* H_0. The investigator must devise a means of generating data for such a test, and then define some quantity, called a *test statistic* $d(\mathbf{X})$, that is a function of the data $\mathbf{X}$ and has a known probability distribution supposing that hypothesis is true. The test statistic should be chosen so that it defines a relevant direction of departure from the null hypothesis. The test statistic should also be defined such that larger values indicate stronger evidence of departure from what is expected if the null hypothesis is true. The probability distribution of the test statistic under the null hypothesis is the *null distribution*. The null distribution thus serves as a mathematical *model* of the null hypothesis, and the direct target of the test is the *statistical hypothesis* that the data are generated by a process characterized by the null distribution. One may then use the null distribution and the observed value of the test statistic to answer the following question*: how probable is it that one would get a value as great or greater than that observed value, assuming the statistical null hypothesis is true*? To the extent that the null distribution is an adequate model of the substantive null hypothesis under the conditions in which the data is generated, the answer to that question will serve as a good estimate of the corresponding probability with regard to the null hypothesis itself. The value of this probability is the *p-value* of the outcome of the test.

To test a null hypothesis H_0 (such as `there is no Higgs boson'), physicists rely on a physical signature of the phenomenon sought after, based on its hypothetical features. Such a signature might come in the form of the decay of a hypothetical particle into other particles identifiable via their measurable properties. Experimenters must then operationalize that physical signature in terms of data selection criteria (*cuts*) that define *candidates* for the phenomenon in question. For a given set of cuts, they must then estimate the rate at which background processes will yield events satisfying those cuts, thus determining the null distribution.

Once the data $\mathbf{x}_0$ are in hand, the observed value of the test statistic $d(\mathbf{x}_0)$ can be recorded and the p-value $\Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0); H_0)$ can be calculated. It has become standard practice in HEP to convert this probability number into a number of σ's by determining what number of standard deviations from the mean of the Standard Normal distribution would correspond to the *p*-value in question.

Introduced by Ronald A. Fisher, significance testing differs from the approach to testing pioneered by Jerzy Neyman and Egon Pearson, though the two approaches share some central concepts and in practice are not kept entirely distinct. In the Neyman-Pearson (NP) approach, both the null hypothesis and the *alternative* hypothesis against which the null is being tested are specified explicitly. The dichotomy between null and alternative hypotheses necessitates the introduction of a corresponding distinction between two types of error. *Type I error* consists of rejecting the null hypothesis when it is true, while *type II error* consists of failing to reject the null hypothesis when it is false. To specify an NP test one chooses first the greatest probability α of committing a type I error that one is willing to allow. This is the *size* of the test. For an NP test with size α, one may then determine, for each element of the alternative hypothesis, the probability (β) of committing a type II error. The *power* of the test, for that element of the alternative, is then defined as $1-\beta$.

The NP framework allows investigators to optimize their tests in the sense that, for a given α, one can specify a test that maximizes power (minimizes β). But NP tests involve tradeoffs insofar as reducing α tends, all else being equal, to increase β, and vice versa.

For reasons that we will not belabor here, HEP practice combines elements of both Fisherian and NP testing approaches (see Staley 2015b). Although the *p*-value to which the 5σ standard is directly applied is a feature of the Fisherian approach, physicists do consider alternative hypotheses and type II error probabilities, as in the NP approach. The crucial point remains that requiring a smaller *p*-value (i.e., more σ's) for a decision to reject the null hypothesis reduces the probability of rejecting the null, assuming it to be correct (type I error), while increasing the probability of failing to reject the null, when it is false (type II error).

These error probabilities provide us with the means to clarify the AIR. Specifying an NP test requires *as an input* a decision about the maximum acceptable type I error rate. Statistical considerations can help the investigator to quantify the trade-off between type I and type II error probabilities. Just how to strike that balance, however, does not follow from any precepts of the statistical theory or from the data. To the extent that the investigator considers non-epistemic values relevant to the acceptable risk of a type I or type II error, they will form the basis of decisions necessary to the specification of an NP test.

3. The AIR and Churchman's project

The AIR, as discussed in the contemporary literature, seeks to establish that considerations of the costs of errors introduce non-epistemic values into scientific reasoning. Here is one possible reconstruction of the argument:

1. Whether given data lead one to accept or reject a hypothesis depends on the inference method chosen.
2. The probability of accepting (rejecting) a hypothesis erroneously depends on the choice of method.

3. Erroneously accepting (rejecting) a hypothesis may have consequences subject to evaluation by non-epistemic criteria.
4. From (3): Non-epistemic values may legitimately influence one's preference between possible inferential errors.
5. From (2) and (4): Non-epistemic values may legitimately influence the scientist's choice of inference methods.
6. From (1) and (5): non-epistemic values may legitimately make a difference to the conclusion the scientist draws from data.

Under this reconstruction, the conclusion only states that non-epistemic values *may* have a legitimate role, but nothing is stated about the scope of their role. Nor does the argument explain what distinguishes epistemic from non-epistemic values.

Discussions of the argument from inductive risk often cite, as the wellspring of our current understanding of the argument, Richard Rudner's (1953) "The scientist qua scientist makes value judgments." Sometimes that citation is accompanied by a citation of C. West Churchman's (1948) "Statistics, pragmatics, induction."[3] Even when Churchman's paper is cited alongside Rudner's the discussion typically focuses on Rudner's paper rather than Churchman's. It is not hard to understand why. Rudner's concise and elegantly written six-page essay is a model of efficient erudition, while Churchman's essay occupies twenty densely-written pages and requires some patience with formalism.

[3] JSTOR lists 41 articles that cite Rudner's essay, but only 8 that cite Churchman's. Of those 8, only three were published later than 1968, beginning with Heather Douglas's (2000), which seems to have brought Churchman's essay back into the discussion.

Churchman's essay is the tip of an iceberg, a précis of Churchman's substantial theorizing about the possibilities of a comprehensive science of ethics as well as the ethics of science. I propose that attending to Churchman's more comprehensive pragmatic approach to scientific inference reveals a way of re-thinking the AIR, focused less on a problematic distinction between epistemic and non-epistemic values, and more on the nature of inquiry itself and the demands of its responsible pursuit.

Churchman begins with a "reformulation," in a language drawn from formal systems theory, of Wald's decision theoretic extension of the NP approach to statistical inference (Wald 1942). Wald defines a "best" decision function for a given statistical problem in terms of the minimization of a risk function that takes into account the "relative importance" of different possible errors.

Churchman regards Wald's account as incomplete: "unless we can formulate the conditions under which a procedure *satisfies specific purposes*, and *is reasonable*, we have left the entire theory of inference in a very unsatisfactory and incomplete status" (Churchman 1948, 254; emphasis added). In response, Churchman develops the *theory of pragmatic inference* and the *theory of induction*. These address, respectively, the question of the ability of an inferential procedure to "satisfy specific purposes," and the question of whether those purposes and the means used to address them are "reasonable."

Churchman notes that for the pragmatist, a statistical inference is a "*means* for accomplishing effectively one or more *ends*" (ibid., 256, emphasis in original), and that pragmatic inference is concerned with finding the most efficient means for achieving any given end. For the purposes of pragmatic inference, then, statistical inference is incomplete. First, it does not describe the procedure by which the data should be acquired or the selection of the presuppositions that are necessary for statistical inference. Second, it does not address the unavoidable question of how acceptance of a hypothesis will be translated into action:

> In pragmatic methodology, every scientific hypothesis is considered to be a possible course of action for accomplishing a certain end, or set of ends. Pragmatically speaking, an inability to say what one intends to do as a result of accepting one out of a set of alternative hypotheses, is an inability to state the hypotheses themselves in adequate terms. Statements like "we merely want to find out so-and-so" represent pragmatically incomplete formulations. (ibid., 259)

Churchman's project aims to formalize pragmatic inference, understood as the problem of selecting a method that reliably chooses --- in a given environment, and on the basis of given data --- that "behavior pattern" (i.e., pragmatically articulated hypothesis) that most efficiently achieves a stated objective. The ideal method would never choose a behavior pattern that was not maximally efficient toward that objective, and "evaluation of a method will depend on departures from the ideal relationship" (ibid., 259).

Pragmatic inference thus understood would suffice, were we able to maximize the efficiency of our pursuit of one end at no cost to the efficiency with which we pursue any other. But clearly this is not so: "in a 'complex' pragmatic situation, we must balance the effectiveness of a method against its inefficiency for certain conflicting ends" (ibid., 261). To deal with this problem we must go beyond pragmatic inference to a theory of *induction*, a term that Churchman uses to refer to the general method of science. An adequate theory of induction would allow us not only to choose the most efficient method for making inferences in the pursuit of any given end, but would provide us with the resources to assign, on the basis of a "science of ethics," relative weights to the ends themselves.

Churchman thus portrays scientific inferences as actions that are always susceptible to two kinds of criticism: as being ill-suited for the aims for which they were chosen, and as being chosen for aims that are inappropriate. Churchman concludes that "*No fact or law of science can be determined without presupposing ethical principles*," but in turn the "proper formulation" of such principles "*depends upon the contributions of the special fields of science*" (Churchman 1948, 266; emphasis in original). This naturalistic ethical investigation will involve a historical inquiry aimed at unearthing human ideals from the development of human societies. The details and feasibility of this project need not concern us here. We can simply note how Churchman concludes that "It is the responsibility of all phases of research, and most particularly statistical research, to become conscious of the need for a science of value which can make explicit contributions to the advance of science, free of vague intuitions about 'reasonableness'" (ibid.).

4. Levi's critique: epistemic autonomy

With Churchman's development of inferential statistics and decision theory more clearly in view, we are in a better position to appreciate Isaac Levi's response, leading to a clearer view of what is at stake in these early discussions of the AIR.

Levi shares with Churchman and Rudner a pragmatic view of scientific inquiry that regards it as involving the *acceptance* of hypotheses. He seeks to defend the view that scientific inquiry aims to replace doubt by true belief. In his 1962 paper "On the Seriousness of Mistakes" he considers two lines of thought that oppose this view. The first denies that scientists accept or reject hypotheses; instead they assign degrees of confirmation to them. The second agrees that scientists do accept and reject hypotheses, but only "in a behavioral sense," and finds support for this interpretation of scientific inquiry in "modern statistical theory" (Levi 1962, 47). Levi attributes the latter view to Churchman and Rudner, and here the focus will be on his argument against it.

According to Levi, a behavioral understanding of acceptance opposes the view that scientists attempt to replace doubt by belief, because it interprets 'accept *H*' as 'acting on the basis of *H*'. The actions involved in such acceptance will vary according to the context in which the decision whether to accept *H* must be made, depending on the objectives the investigator seeks to realize. Consequently, "the 'theoretical' objective of accepting only true propositions as true is hopelessly ambiguous" (ibid., 48).

Levi begins his countervailing view by distinguishing two types of attempts to replace doubt by true belief.

Type (a) inferences are attempts to "seek the truth and nothing but the truth." The scientist must select the true proposition from a set of competing possible propositions on the basis of the relevant evidence. Two constraints are operative in type (a) inferences, according to Levi: (1: *hypothesis impartiality*) The investigator must not prefer that any one of the propositions be true rather than another. (2: *error impartiality*) She also must regard each possible mistake as being equally serious.

In type (b) inferences, one seeks to replace doubt with a belief that possesses "certain desirable properties *in addition to* truth" (ibid., 49; emphasis in original), such as simplicity or explanatory power. In such inferences, the investigator is not obligated to be impartial as to which proposition is true, and may also conduct inquiry in a way that treats some errors as worse than others.

The denial that science relies on type (a) inferences puts significant pressure on a non-behaviorist view, leading to a worry "that the notion of accepting a hypothesis in a non-behavioral sense might be entirely dispensable in an account of inductive inference" (ibid., 51).

Levi's response to this state of affairs is to defend the relevance of type (a) inquiries to science. At issue is the feasibility of carrying out inquiry under the constraints that apply to such inferences, particularly error impartiality, the viability of which the AIR clearly targets. His arguments respond to the claim that inferences carried out within the framework of modern statistical theory require a distinction between contrasting types of errors, and that determining standards for the acceptance of hypotheses under consideration cannot be accomplished without at least implicit consideration of the differential costs of those errors.

In short, what Levi proposes is a reinterpretation of the NP approach to testing that seeks to eliminate the differential treatment of type I and type II errors, and thus render NP compatible with error impartiality. To achieve this, Levi proposes that outcomes that fall outside the *critical region* (the region of test statistic values that result in rejection of the null hypothesis) should lead to *suspension of judgment* rather than acceptance of the null hypothesis. Under this proposal, the distinction between type I and type II errors

> is not a distinction between two different kinds of mistakes (in the sense that a mistake results when a false hypothesis is accepted as true) but between the result of rejecting the null hypothesis when it is true (type I error) which is a bona fide mistake and the result of suspending judgment when the null hypothesis is false (type II "error") which is not. Consequently, type II error can be said to be less serious than type I error without violating the requirement that a person seeking the truth and nothing but the truth take all mistakes with equal seriousness. (ibid., 62--63)

The level of significance itself, of course, remains a matter of choice on the part of the investigator and Levi proposes that on his account, this serves as "a rough index of the degree of caution exercised in a search for truth" (ibid., 63).

Levi's interpretation of the outcomes of significance tests in terms of a choice between rejecting the null and suspending belief regarding it is certainly plausible and reflects actual practice to a significant degree. What are its implications for the AIR?

By reformulating significance test inferences so that only one mistake is possible (rejecting the null hypothesis when it is true), Levi negates the premise that scientists value differently the negative consequences of different mistakes.

Scientists, however, may regard not only errors differently but outcomes in general. A suspension of judgment is an outcome no less than an acceptance, and it is perfectly reasonable, on non-epistemic grounds, to have a different attitude towards that outcome than one might have towards outcomes that involve rejecting the null hypothesis, correctly or incorrectly.

We can therefore reformulate the argument introduced in section 2 as follows:

1'. Whether given data lead one to accept, reject, or suspend judgment regarding a hypothesis depends on the inference method chosen.

2'. The probability of any particular inferential outcome depends on the choice of method for drawing inferences.

3'. The outcomes of inferences may have consequences subject to evaluation by non-epistemic criteria.

4'. From (3'): Non-epistemic values may legitimately influence one's preference between possible inferential outcomes.

5. From (2') and (4'): Non-epistemic values may legitimately influence the scientist's choice of inference methods.

6'. From (1') and (5): Non-epistemic values may legitimately make a difference to the outcome of an inference from data.

Where does this leave Levi's category of type (a) investigations: attempts to "seek the truth and nothing but the truth"? Recall that Levi's ultimate concern is with a behaviorist view that regards 'accepting *H*' as equivalent to 'acting on the basis of *H*' as called for by the decision problem at hand. Levi's response in his 1962 paper to such behaviorism is to defend the claim that there are scientific inquiries that are appropriately understood as type (a) inquiries.

How should we conceive of such inquiries? In his (1967) Levi argues that one may fruitfully understand inductive reasoning via a decision theoretic approach without collapsing into behaviorism, thus revealing a strong analogy between practical decision problems and cognitive decision problems that does not reduce the latter to the former.

Using a Bayesian framework, Levi recasts type (a) inquiries as *cognitive decision problems* employing both probabilities and *epistemic utilities*.[4] An epistemic utility function is a means for ordering possible inferential outcomes with respect to epistemic preference. Levi imposes two conditions on such orderings that replace conditions (1) and (2) in his 1962 paper: "(1') Correct answers ought to be epistemically preferred to errors. (2') Correct answers (errors) that afford a high degree of relief from agnosticism ought to be preferred to correct answers (errors) that afford a low degree of relief from agnosticism" (Levi 1967, 76; notation altered for clarity).

The application of condition (2') is facilitated by Levi's introduction of a measure of relief from agnosticism, based on logical relations among sentences considered as relevant possible answers to the question investigated. I will not discuss the details of Levi's account, but will focus instead on the roles played by Levi's cognitive decision problems and practical decision problems in the production of scientific knowledge.

First, we should note that we can use Levi's conditions (1') and (2') to relieve an explicative obligation that has burdened the discussion thus far: what is an epistemic value? We need not answer this question, because we can replace the terminology of epistemic values and judgments about them with the terminology of epistemic utility functions, which are exactly those utility functions that satisfy conditions (1') and (2'). Any utility function that violates these two conditions is therefore non-epistemic, whatever the value judgments that generate them.

[4] Although Levi's framework uses Bayes' rule for decision purposes, his solutions to cognitive decision problems are not Bayesian inferences in the usual sense, as they result in decisions whether to accept a hypothesis rather than in the determination of posterior probability functions.

It would seem that we could reconcile the AIR as formulated here with Levi's defense of type (a) investigations by exploiting the vagueness of the term 'influence' in premise 4. Levi incorporates parameters for the "degree of caution" in an inference both in his 1962 paper (significance level) and in his 1967 book ("q-index"). With regard to the latter he allows that the choice of a q-index, which is necessary for defining an epistemic utility function, "is a subjective factor which does in some sense reflect the investigator's attitudes" (Levi 1967, 89). What is important regarding the q-index is not the considerations on which it is based (the values that "influence" that choice), but that the choice of a q-index value constitutes a "commitment on the part of an investigator to have his conclusions evaluated according to certain standards" (ibid.).

The resulting view seems rather close to the view expressed by Heather Douglas (2009). According to Douglas' critique of the value-free ideal in science, the important distinction is not between epistemic and non-epistemic values, but between different roles that value judgments might play in different aspects of scientific inquiry. In their direct role values "act as reasons in themselves to accept a claim," whereas in their indirect role they "act to weigh the importance of uncertainty about the claim, helping to decide what should count as *sufficient* evidence for the claim" (Douglas 2009, 96; emphasis in original). Value judgments, according to Douglas, can play an indirect role throughout scientific reasoning, but must not play a direct role in the "later stages" of scientific inquiry, during which the scientist decides how to interpret data, how much support data lend to competing hypotheses, and whether to accept or reject hypotheses under investigation.

Douglas understands Levi to be a defender, even the chief proponent, of the value free ideal of science that she criticizes (ibid., 90). Our current perspective on Levi's views reveals a strong affinity with Douglas' own position. On Levi's account, value judgments of many kinds might (in their "indirect role"?) contribute to a decision to let one's inferences be governed by a determinate degree of caution. But the autonomy of the epistemic is preserved insofar as the inferences thus carried out can be understood strictly as attempts to replace agnosticism with belief, such that correct answers are preferred to incorrect answers, and such that the decision whether to accept a hypothesis as an answer rather than suspend judgment is based on a balance between an interest in relieving agnosticism and a cautiousness regarding the risk of embracing an answer erroneously. A threat to the autonomy of the epistemic would arise from employing in a cognitive decision problem a utility function that gave preference to some hypotheses over others in violation of constraints (1') and (2'). Any value judgment that gave rise to such a utility function could indeed be thought of as supplanting the role of evidence, which is precisely what Douglas seeks to rule out with her prohibition of values playing a direct role in the later stages of inquiry.

To be sure, there remains more to be said about the relationship between Douglas's account and Levi's, but I hope this suffices to establish an interesting question that deserves further consideration in another setting.

To return to the main thread of this essay, let us grant that Levi has established the possibility of carrying out type (a) inquiries that seek "the truth and nothing but the truth." We are left with the question of the relevance of such inferences to the practice of science. Does conceiving of scientific inquiry in terms of seeking "the truth and nothing but the truth" suffice for understanding how scientific knowledge is generated out of scientific inquiry?

I do not propose to be able to definitively answer this question here, but I do aim to use the example of the Higgs search at the Large Hadron Collider to suggest some reasons for thinking a negative answer to this question plausible.

5. The 5σ standard

Many of the reports in the news media that followed the Higgs announcements by CMS and ATLAS focused on their claims of having results that were significant at a level of 5σ, a level that was presented as a "gold standard in physics for discovery" in the *New York Times* and as codifying "strict notions of scientific certainty" in the *Washington Post* (Overbye 2012; Vastag and Achenbach 2012).

The actual role of the 5σ standard in HEP is not easily characterized. It has no official status as a rule by which HEP investigators are bound, and physicists will deny that its normative force is absolute. Joe Incandela, who served as spokesperson for CMS at the time of the Higgs announcements, has stated that "the 5 sigma standard is generally misunderstood outside the field. We do not take 5 sigma as absolutely necessary nor do we assume all 5 sigma results to be correct" (personal communication). Similarly, CMS member Robert Cousins comments, "I do not believe that experienced physicists have such an automatic response to a *p*-value, but it may be that some people in the field may take the fixed threshold more seriously than is warranted" (Cousins 2014, 30).

Although the 5σ standard has acquired some degree of importance as a guidepost in deliberations over how to present new findings in physics, its rationale escapes any simplistic codification.[5] Here I will limit my discussion to factors cited as relevant to its application in a recent essay by physicist Louis Lyons aimed at reforming the use of the 5σ standard.[6]

While advocating the adoption of more variable discovery standards, Lyons articulates four factors that should be relevant to the determination of such standards (Lyons 2013). He proposes a scheme (in which the criterion for the Higgs discovery remains at 5σ) according to which the discovery standard for a given potential finding is based on: (1) the *degree of surprise* (which Lyons also calls the "sub-conscious Bayes' factor"), (2) the *impact* of the discovery, (3) the size and salience of *systematic uncertainties* in the search procedures, and (4) the presence of a *Look Elsewhere Effect* (in other contexts called a "multiple trials" effect). Each of these considerations has been cited elsewhere in comments by HEP physicists explaining the rationale for the 5σ standard. Here the focus will be on (2): *impact*.[7]

Although Lyons includes impact among the relevant factors he does not expand on how this factor should be understood. For the case of the Higgs search, however, some impacts are readily discernible from consideration of the current state of HEP, while others come to light in comments from physicists themselves. Values implicated in these

---

5 The story of how the 5σ standard acquired that status has recently been documented by Allan Franklin, who notes that the 5σ standard has had a gestation period extending across twentieth-century particle physics experimentation, but has only recently acquired its present status of a presumptive standard (Franklin 2013).

6 Although this limitation reflects a decision not to attempt a survey of the views of physicists, the factors that Lyons cites cohere well with views expressed via internet postings by physicists, personal communications with members of ATLAS and CMS, and the writings of physicists familiar with the issues.

7 Dawid (2015a; 2015b) discusses (1) and (3); for further discussion of (3) see Cousins 2014; Gross and Vitells 2010; Lyons 2013; Staley 2015b; Vitells 2011; for (4) see Cousins and Highland 1992; Mari and Giordani 2014; Staley 2015a.

outcomes relate to the value of the discovery claim itself, as well as to the potential harms caused by making a discovery claim that turns out to be erroneous.

I will focus my attention on impacts that fall into two broad categories: those relating directly to argumentation in future physics inquiries and those indirectly related to the broader goals of the ATLAS and CMS groups, the HEP community, and scientists generally.

Regarding the first category, accepting the existence of a new boson involves a commitment (or at least a license) to adopt statements entailing the existence of such a particle as premises in the pursuit of further inquiries. Such a commitment affects the continued work of ATLAS and CMS, as their analytic tasks turn from the aim of producing exclusion plots (showing, for example, what hypothetical Higgs masses have been ruled out) towards the aim of measuring the properties of the newly discovered particle to fix more securely the theoretical interpretation of their finding (such as whether it is truly a Standard Model Higgs boson). For other physicists working on SM and Beyond-SM problems, the announcement by ATLAS and CMS has the consequence of changing the logical terrain. Although each investigator must decide (as an individual or as a member of a group) whether the evidence offered by the two CERN groups suffices to warrant accepting the existence of the Higgs as a premise or assumption in future work, it seems likely that the burden now lies on those who would decline those claims to explain their dissent. These considerations contribute to our understanding of the 5σ standard for the Higgs search by highlighting the importance, for the pursuit of physics inquiries, of guarding against an erroneous discovery claim, while also pointing towards the tremendous value of that discovery claim, as it enables the pursuit of new inquiries that, prior to discovery, had to wait offstage.

Judging what belongs to the second category of consequences calls for a more speculative approach, but various statements of physicists involved in the Higgs search provide some clues. CMS's published paper declares in its introduction that "The discovery or exclusion of the SM Higgs is one of the primary scientific goals of the Large Hadron Collider" (Chatrchyan et al. 2012, 30). Given the great expense of building the LHC and operating the CMS and ATLAS experimental programs, it is not surprising that success at achieving this goal was highly valued. The much-anticipated discovery claims were articulated not only in detailed scientific talks aimed at the physics community but also in a presentation to the media that was broadcast via the internet worldwide and featured prominently among the news of the day. To get things wrong would have been tremendously embarrassing. A comprehensive assessment of all risks of such an error would include a political dimension with potentially negative consequences for the funding of HEP.

In addition to concerns about the amount of effort and expense that had gone into the search for the Higgs and its importance to the scientific project of the LHC, a broader sense of responsibility toward the public perception of science in general may have played a role in the cautious attitude toward any discovery announcement. According to CMS member Robert Cousins, the intense public spotlight that the LHC had felt since 2008 made it clear that there was an opportunity to try to show science of very high quality to the general public, in an environment where there was public skepticism about some scientific claims. Certainly, making a discovery announcement that subsequently turned out to be erroneous carried a very high cost, and could only contribute to such skepticism (personal communication).

Whether they are "internal to science" or pertain to broader goals, the impacts just mentioned include both the costs of error and the benefits of getting it right. Are these epistemic considerations? Daniel Steel has proposed that what is distinctive of epistemic values is that they "promote the attainment of truth," either *intrinsically*, in that "manifesting that value constitutes an attainment of or is necessary for truth" or *extrinsically*, in that manifesting the value promotes the attainment of truth without constituting the attainment of or being necessary for truth (Steel 2010).

Relying on Steel's proposed criterion, the impacts here discussed seem to qualify as epistemic, but only in the extrinsic sense. The benefits of correctly accepting the claim to have discovered a new boson in no way make that claim more likely to be true. Neither are they prerequisites for its being true. Similarly, the claim of a newly discovered boson is not made more likely to be true, or better supported by evidence, by the fact that, were it in error, other investigations relying on that claim as a premise would also lead to errors, much less by the embarrassment of particle physicists or their loss of prestige or funding.

Levi's critique, however, casts doubt on the importance of classifying values as epistemic or non-epistemic. For Levi preserving the epistemic autonomy of science is not a matter of excluding judgments about non-epistemic values (whatever those are) from exerting an influence on scientists' reasoning, but of securing within science a mode of reasoning in which no preference is given to the acceptance of any among the competing hypotheses. Levi contends that such epistemic autonomy is preserved so long as whatever value judgments the scientist makes exert their influence *only* via the determination of an investigator's degree of caution.

Framing the question in *this* way refocuses our attention away from a distinction regarding values and toward a question about the *kind* of inference problem at hand: For *what* is the 5σ standard a standard?

It should be clear by now that it is not a standard to be used strictly in one of Levi's cognitive decision problems; its application in the Higgs case is not to an attempt to seek the truth and nothing but the truth. Rather, it is a standard governing the decision of *how to report* the outcome of the experimental search for the Higgs boson. This decision concerns not only cognitive but also communicative actions. Indeed, it is hard to see how any inference drawn by an investigator can possibly constitute a contribution to science without that investigator making such a decision. In the (pervasive) context of a collective investigative undertaking, the formation of individual belief is especially remote from the decisions that produce usable results, as those decisions involve distributed deliberative procedures.

One might, of course, insist that nonetheless each individual scientist might (or even must), before deciding how to communicate a result, decide what to believe regarding that result, making the cognitive decision prior to the communicative decision. However, even supposing this to be the case, the influence of such cognitive decisions will remain *mediated* through the communicative decisions that follow upon them. Allowing, therefore, that Levi's cognitive decisions occur as the outcome of attempts to seek the truth and nothing but the truth, they are insufficient for the production of scientific knowledge. For *that* investigators must decide on the most *beneficial* action to take in response to the results in hand. It may well be that, as in the Higgs search, the benefits that are relevant to that decision are not practical in the narrow sense, but accrue to science understood as a knowledge-generating enterprise. Nonetheless, it is a practical decision --- i.e., one regarding not just what to believe but what action to take.

This conclusion might seem entirely friendly to Levi's aims: We might frame the practical decision at hand simply in terms of choosing which, among the hypotheses under consideration, to report as that supported by the outcomes of the experiment. Suppose, then, that the utility function used in this simple decision problem *preserves* the ordering of the epistemic utility function used in the underlying cognitive decision problem. In this way we might preserve epistemic autonomy in Levi's sense. All that would have changed is that, whereas Levi's cognitive decision problems are merely analogous to practical decision problems, this class of decision problems, lying at the very heart of scientific knowledge generation, would *be* practical decision problems.

And that is the problem for epistemic autonomy. As a practical problem, the decision about communicating the outcome of an experiment is subject to the full range of utility considerations applicable to any practical decision, even if considered in the simple terms just suggested. One might, of course, decide in the end that such a decision should respect the ordering imposed by an epistemic utility function. Upholding such a restriction, however, would be the outcome of a consideration of broader utilities and could not be guaranteed at the outset.

Moreover, such decisions are not simple in the manner just suggested. The communication of scientific results is an outcome of a much more complicated decision that must also consider, among other things, *how* to report the results. For example, in a search for a new phenomenon in HEP, should one claim "evidence for" or "observation of" the phenomenon? This is one sense in which framing such decisions as purely a matter of what to believe remains, in Churchman's words, "pragmatically incomplete."

6. Conclusion

Let me conclude with a clarification of the position staked out in this paper.

Recall that the concern that motivated Levi's critique was not to refute the AIR as that has come to be understood, but to refute a reductive behaviorism regarding statistical inference that he associated with the work of statisticians like Jerzy Neyman as well the philosophical work of Rudner and (especially) Churchman. When I claim that Levi's defense of the epistemic autonomy of scientific inference has at most limited scope, I do not thereby intend to defend a reductively behaviorist understanding of statistical inference.[8] On the contrary, the error probabilities of frequentist statistical procedures such as significance testing or NP testing are guides to the planning of efficient and reliable strategies for data collection and interpretation, and contribute (along with additional considerations, such as severity analysis (Mayo and Spanos 2006)) to the post-data determination of what inferences may justifiably be drawn from the data. However, the drawing and reporting of inferences in the context of scientific inquiry is not merely a

---

[8] Nor do I endorse Levi's reading of the mentioned authors as advocating reductive behaviorism. Neyman does advocate what he himself calls "inductive behaviorism," but one should also consider Neyman's more nuanced statements (e.g. Neyman 1976), as Deborah Mayo and Aris Spanos have pointed out (Mayo and Spanos 2006). I regard Churchman's behaviorism as better understood not in terms of reductive behaviorism but in terms of a Peircean pragmaticist project of achieving a higher grade of clarity. Rudner is implicated in behaviorism only insofar as he invokes the statistical methods of Neyman and Pearson, which are not inherently behavioristic. Another essay would be required to flesh out and defend this stance.

matter of forming beliefs. It is instead a practical matter and as such is open to the full range of value considerations that bear on our decisions in every domain of activity. If we choose to build scientific knowledge in a way that preserves epistemic autonomy, it will be because we think doing so will deliver the greatest good, all things considered.

References

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., . . . Zwalinski, L. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters*, B716, 1–29. doi: 10.1016/j.physletb.2012.08.020.

Allen, R. (2014). The Higgs bridge. *Physica Scripta*, 89, 018001.

Chatrchyan, S., Khachatryan, V., Sirunyan, A., Tumasyan, A., Adam, W., Bergauer, T., . . . Swanson, J. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters*, B716, 30–61. doi: 10.1016/j.physletb.2012.08.021.

Churchman, C. W. (1948). Statistics, pragmatics, induction. *Philosophy of Science*, 15(3), 249–268.

Cousins, R. D. (2014). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, 1–38. Retrieved from http://dx.doi.org/10.1007/s11229-014-0525-z doi: 10.1007/s11229-014-0525-z.

Cousins, R. D., & Highland, V. L. (1992). Incorporating systematic uncertainties into an upper limit. *Nuclear Instruments and Methods in Physics Research*, A320, 331–335.

Dawid, R. (2015a). Bayesian perspectives on the discovery of the Higgs particle.

Synthese, 1–18. Retrieved from http://dx.doi.org/10.1007/s11229-015-0943-6 doi: 10.1007/s11229-015-0943-6.

Dawid, R. (2015b). Higgs discovery and the look elsewhere effect. *Philosophy of Science*, 82(1), 76–96.

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, 67(4), pp. 559-579.

Douglas, H. (2009). *Science, policy, and the value-free ideal*. Pittsburgh: University of Pittsburgh Press.

Franklin, A. (2013). *Shifting standards: Experiments in particle physics in the twentieth century*. Pittsburgh, PA: University of Pittsburgh Press.

Gross, E., & Vitells, O. (2010). Trial factors for the look elsewhere effect in high energy physics. *European Physical Journal* C, 70(1-2), 525–530.

Levi, I. (1962). On the seriousness of mistakes. *Philosophy of Science*, 29(1), 47–65.

Levi, I. (1967). *Gambling with truth: An essay on induction and the aims of science*. New York: Alfred A. Knopf.

Lyons, L. (2013). Discovering the significance of 5σ. (arXiv:1310.128).

Mari, L., & Giordani, A. (2014). Modelling measurement: Error and uncertainty. In M. Boumans, G. Hon, & A. Petersen (Eds.), *Error and uncertainty in scientific practice* (pp. 79–96). London: Pickering and Chatto.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman-Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, 57(2), 323–357.

Myrvold, W. C. (2012). Epistemic values and the value of learning. *Synthese*, 187, 547–

568.

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in Statistics – Theory and Methods*, A5, 737–751.

Overbye, D. (2012, July 4). Physicists find elusive particle seen as key to universe. *New York Times*.

Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1), 1–6.

Staley, K. W. (2015a). Estimation of systematic uncertainty as robustness analysis. (Unpublished).

Staley, K. W. (2015b). Pragmatic warrant for frequentist statistical practice: The case of high energy physics. Retrieved from http://philsci-archive.pitt.edu/10769/ (Unpublished).

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, 77(1), 14–34.

Vastag, B., & Achenbach, J. (2012, July 4). Scientists' search for Higgs boson yields new subatomic particle. *The Washingon Post*.

Vitells, O. (2011). Estimating the "look elsewhere effect" when searching for a signal. In H. B. Prosper & L. Lyons (Eds.), *Proceedings of the PHYSTAT 2011 workshop on statistical issues related to discovery claims in search experiments and unfolding* (pp. 183–189). Geneva: CERN.

Wald, A. (1942). *On the principles of statistical inference: Four lectures delivered at the University of Notre Dame*. Notre Dame, Indiana: University of Notre Dame Press.