Synthese (2017) 194:355–376
DOI 10.1007/s11229-016-1111-3

S.I. : EVIDENCE FOR THE HIGGS PARTICLE

Pragmatic warrant for frequentist statistical practice: the case of high energy physics

Kent W. Staley[1]

Received: 13 June 2014 / Accepted: 29 April 2016 / Published online: 11 May 2016

Abstract Amidst long-running debates within the field, high energy physics (HEP) has adopted a statistical methodology that primarily employs standard frequentist techniques such as significance testing and confidence interval estimation, but incorporates Bayesian methods for limited purposes. The discovery of the Higgs boson has drawn increased attention to the statistical methods employed within HEP. Here I argue that the warrant for the practice in HEP of relying primarily on frequentist methods can best be understood as *pragmatic*, in the sense that statistical methods are chosen for their suitability for the practical demands of experimental HEP, rather than reflecting a commitment to a particular epistemic framework. In particular, I argue that understanding the statistical methodology of HEP through the perspective of pragmatism clarifies the role of and rationale for significance testing in the search for new phenomena such as the Higgs boson.

Keywords Higgs boson · Statistics · Significance testing · p values · Frequentism · Pragmatism

1 Introduction

On July 4, 2012 the CMS and ATLAS collaborations announced their latest findings from the search for the Higgs boson. ATLAS spokesperson Fabiola Gianotti declared that they had observed "clear signs of a new particle, at the level of 5 sigma, in the mass region around 126 GeV" (ATLAS 2012). The CMS statement reported the observation of an "excess of events at a mass of approximately 125 GeV with a

✉ Kent W. Staley
staleykw@gmail.com

[1] Saint Louis University, St. Louis, USA

statistical significance of five standard deviations above background expectations …. We interpret this to be due to the production of a previously unobserved particle with a mass of around 125 GeV" (CMS 2012). CMS and ATLAS considered the evidence insufficient to declare that the new particle was the Higgs boson itself, but only stated that the evidence was consistent with the expectations from decays of a Higgs boson.

Press coverage emphasized the appeal in these declarations to a standard of discovery: in order to announce the discovery of a new particle, the physicists needed to show that they had found an excess of candidate events beyond the expectations from background alone that would constitute a departure of at least five standard deviations ("5σ"). The associated probability statement (p value) was reported variously. The New York Times reported that "Both groups said that the likelihood that their signal was a result of a chance fluctuation was less than one chance in 3.5 million, 'five sigma,' which is the gold standard in physics for a discovery" (Overbye 2012). Reuters noted that "Five sigma, a measure of probability reflecting a less than one in a million chance of a fluke in the data, is a widely accepted standard for scientists to agree the particle exists" (Wickham and Evans 2012).

Meanwhile, in discussions on the website of the International Society for Bayesian Analysis, statisticians debated the statistical methodology employed in the Higgs discovery. Tony O'Hagan, prompted by "[a] question from Dennis Lindley," posted a series of queries about the Higgs search results, referring to the 5σ requirement as "extreme" and asking for its justification. O'Hagan stated, "Rather than ad hoc justification of a p value, it is of course better to do a proper Bayesian analysis. Are the particle physics community completely wedded to frequentist analysis?" and asked, "If so, has anyone tried to explain what bad science that is?" (O'Hagan 2012).[1]

These questions put into play two distinct issues regarding the statistical methodology of HEP. One is HEP's reliance on the 5σ standard for discovery claims. The other is the use of the methodology of significance testing. In this paper, I will focus on the question of the warrant for HEP's reliance on significance testing, though this will lead naturally to consideration of the 5σ standard. The second of O'Hagan's questions explicitly assumes that "it is better to do a proper Bayesian analysis." Were this the case, then the use of significance testing in HEP would indeed be puzzling, and one would want to investigate the reasons for the persistent failure of presumably well-trained and mathematically competent scientists to take advantage of the availability of a better method of analysis than that which they use. I will argue that O'Hagan's presupposition is incorrect: the use of significance testing in such contexts as the Higgs search is well-warranted, and a Bayesian analysis is not "of course better." Understanding why requires consideration of the specific and limited purpose for which significance testing was used in the Higgs search.

My intention in this paper is not, however, to trudge along the well-worn paths of the debates between frequentists and Bayesians (though my course might intersect

[1] O'Hagan collected and summarized the many replies he received to his post. In this digest, he noted that he had intentionally used somewhat inflammatory language to "provoke discussion" (O'Hagan 2012).

these at some points).[2] Rather, I will ask what warrants the application of significance testing to the specific tasks for which HEP employs such tests?

I argue that the use of significance testing by scientists pursuing the discovery of the Higgs boson is warranted because of its strategic value in (1) enabling physicists to determine whether their data are, in a relevant way, statistically discrepant from the hypothesis asserting that the decay of Higgs bosons does not contribute to their data, and (2) doing so in a way that enables them to present a cogent argument appropriate to their anticipated audience. Moreover, I argue that their reliance on the 5σ standard for discovery is warranted by their consideration of both the negative consequences of an erroneous discovery claim and the value for the further pursuit of inquiry of a correct discovery claim. These warrants are independent of philosophical commitments regarding the meaning or ontology of probability statements, or the relationship between probability and belief. In this way, I show the warrant for the use of significance testing in HEP to be pragmatic insofar as it is grounded in the practical demands of scientific discovery and argumentation.

I begin my argument in Sect. 2 with a quick summary of the argument given by the ATLAS collaboration in their paper announcing the discovery of a new boson with Higgs-like properties, highlighting the role of their appeal to significance testing. Sect. 3 explains the pragmatic warrant for HEP's use of significance testing. In Sect. 4 I offer some brief comments on the 5σ standard in light of the argument previously given. Section 5 takes up an objection and emphasizes the priority of argumentative tasks over statistical methodology. The paper's conclusions are summarized in Sect. 6.

2 Significance testing in HEP: the case of the Higgs

By characterizing their evidence in terms of an estimate of the statistical significance of their findings, ATLAS and CMS adopted the language and methodology of significance testing, a statistical methodology for testing hypotheses that relies only on probabilities understood as relative frequencies. I will begin with a rough summary of the argument used by ATLAS (the reasoning in the paper submitted by CMS is similar) in order to highlight the role played by appeals to significance testing in ATLAS's experimental argument. Understanding this role is, I claim, essential for understanding the pragmatic warrant for using significance testing.

ATLAS bases its discovery claim on two distinct periods of data-collection. The 2011 dataset was collected with the LHC operating at a center-of-mass energy of $\sqrt{s} = 7$ TeV, while the 2012 dataset came from a $\sqrt{s} = 8$ TeV run. Both ATLAS and CMS had already found excesses (Aad et al. 2012a; Chatrchyan et al. 2012a) beyond background expectations in the 2011 data "compatible with SM Higgs boson production and decay in the mass region 124–126 GeV, with significances of 2.9 and 3.1 standard deviations, respectively" (Aad et al. 2012b, p. 1).

Crucial to the argumentative strategy of the ATLAS paper is the identification of distinctive decay modes of the Higgs boson that lead to distinct search strategies. A

[2] For a well-informed guided tour of those paths, with some novel insights, see Sprenger (2016) who cites the Higgs case as motivation for a careful consideration of the issues.

Standard Model (SM) Higgs boson has a number of distinct decay modes: $\gamma\gamma$, WW, ZZ, $\tau\tau$, bb, $Z\gamma$, and $\mu\mu$, among others. ATLAS's discovery claim rests on data from the first three. Data are selected according to criteria (*cuts*) tailored to the expected features of particles decaying in these modes. The number of such *candidate events* is compared to the number of events expected to satisfy the cuts that result from background—i.e., from processes involving already established physics. An excess number of candidate events beyond what is expected from background might be evidence of the existence of a new boson such as the Higgs, or it might be the result of an upward fluctuation in the rate of background processes.

The statistical significance calculation contributes to a judgment of the plausibility of the latter scenario by determining the probability of getting an excess as large as or larger than that observed, under the supposition that only background processes are involved. That probability is the p value of the observed excess.

The three decay modes from which the evidence is drawn add evidential weight to the statistical significance argument insofar as they help to fix the theoretical interpretation of the excess indicated: the excesses show up in multiple channels in a manner that is predictable in light of knowledge about the rates at which the Higgs should decay in those channels and the size of the backgrounds in each of them.

The guidance of theory is important to the validation of ATLAS's evidence claim in another way. ATLAS relies on statistical models of both the background and the signal for a SM Higgs boson. Neither of these statistical models can be calculated directly from theory. Both require the use of simulation (Massimi and Bhimji 2015; Morrison 2015). For the signal, this simulation does depend, however, on a theoretical characterization of the processes by which Higgs bosons are produced (see, e.g., Harlander and Kilgore 2002). Understanding the signal is important both for developing and optimizing the analytic procedures to be applied to data, and for the comparison of the observed excess with that expected for an SM Higgs with a mass near that reported.

For present purposes, the latter consideration is particularly salient: in addition to demonstrating that they have achieved a statistical significance in excess of 5σ, ATLAS presents a comparison between the excess that they observe and what one would expect from SM Higgs decays near a mass of 125 GeV (see Fig. 1b). Moreover, they do this not only for the combined results, but also separately for the results from each of the three decay channels that figure in their discovery. In each case, it is important that the results fit, at least at a qualitative level, with the expectations from an SM Higgs boson with $m_H \sim 125$ GeV, and not so well with the expectations for a Higgs with mass far from that value. Such an agreement amongst different search modes would not be likely for data generated by background processes alone.

Another important aspect of the ATLAS argument is their characterization of the excess that they find. In particular, they estimate the mass of the new particle that they have observed using the profile likelihood ratio $\lambda(\mu, m_H)$.[3] ATLAS presents a plot

[3] When confronted with a statistical model with multiple parameters, all but one of which are considered 'nuisance' parameters, the profile likelihood for the parameter of interest is obtained by maximizing, for every considered value of the parameter of interest, the likelihoods for the each of the nuisance parameters, and then using the values for the nuisance parameters thus obtained for estimating the parameter of interest (Cox 1970; Venzon and Moolgavkar 1988).

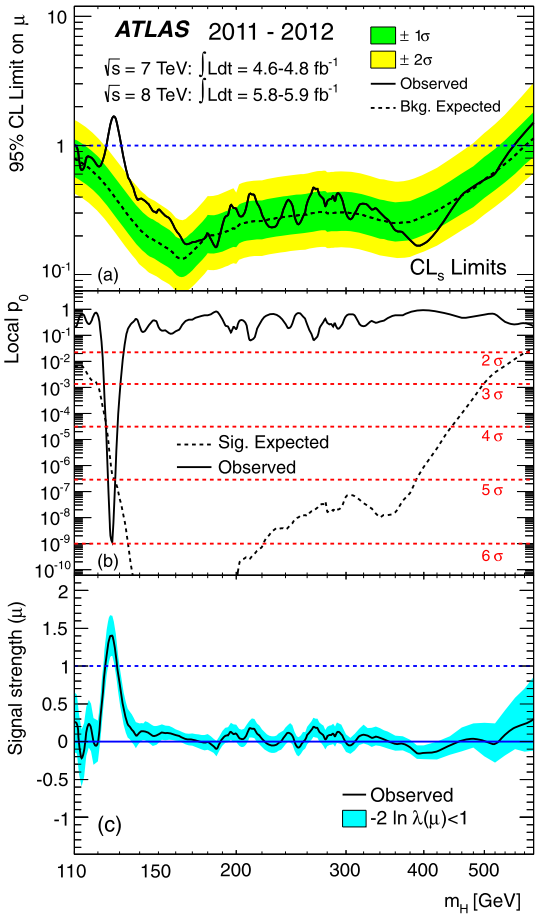

Fig. 1 Three important ways of evaluating the ATLAS results. In **a** the *solid line* indicates 95 % upper bounds on the value of μ established by the observed data, while the *dotted line* indicates the upper bounds that would be expected for background only, with bands showing the $\pm 1\sigma$ and $\pm 2\sigma$ deviations on those background-only upper bounds. In **b** the *solid line* gives the local p value as a function of m_H, while the *dotted line* indicates the expected p value based on simulation of the signal, also as a function of m_H. The best-fit estimate $\hat{\mu}$ of the signal strength as a function of m_H is given in (**c**) (Aad et al. 2012b, p. 13)

that shows the 68 and 95 % confidence intervals in the (μ, m_H) plane for each of the $H \rightarrow \gamma\gamma$, $H \rightarrow ZZ$, and $H \rightarrow WW$ channels. The first two form distinct but overlapping contours, while the latter yields no lower bound on m_H (Fig. 2). ATLAS notes that the difference between the maximum likelihood estimates for m_H based on the $H \rightarrow ZZ$, and $H \rightarrow \gamma\gamma$ channels is sufficiently great that there is only about an 8 % probability "for a single Higgs boson-like particle to produce resonant mass peaks [in those two channels] separated by more than the observed mass difference" (Aad et al. 2012b).

Now, let us consider more closely ATLAS's calculation of a p value for their results. This calculation figures crucially in the way that ATLAS describes the excess of candidate events beyond background expectations.

As a significance test, the hypothesis tested by ATLAS is framed in terms of the value of a parameter μ, called the *signal strength*. This parameter, which acts as a "scale factor on the total number of events" that the SM predicts for the Higgs signal, is a function of m_H, the (unknown) Higgs mass. It is defined so that $\mu = 0$ corresponds

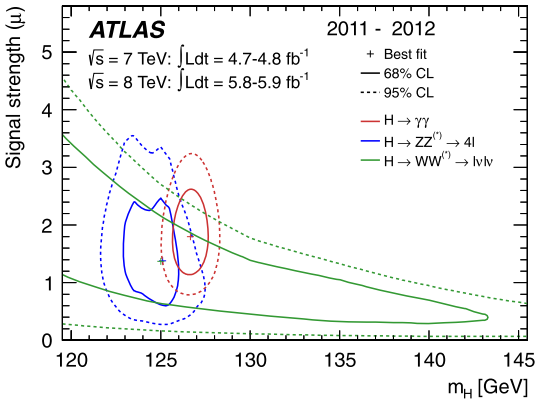

Fig. 2 Confidence intervals in the (μ, m_H) plane for the $H \to \gamma\gamma$, $H \to ZZ$, and $H \to WW$ channels. Maximum likelihood estimates $(\hat{\mu}, \hat{m_H})$ are marked with '*plus*' (Aad et al. 2012b, p. 14)

to the background only hypothesis and $\mu = 1$ corresponds to the hypothesis that an SM Higgs boson as well as background contributes to the data. This allows $\mu = 0$ to serve as the null hypothesis subjected to significance testing. (The dependence of μ on m_H introduces complications to the calculation and interpretation of p values, as discussed below.)

Testing this null hypothesis requires a choice of *test statistic*. The test statistic must be a function of the data $\mathbf{X}$ with a known probability distribution supposing that hypothesis is true (the *null distribution*), and 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, in the relevant direction, from what is expected if the null hypothesis is true.

Determining the null distribution in a search for a new particle amounts to estimating the rate at which background processes will yield events satisfying the cuts. ATLAS follows a common practice in HEP; they do not use simply the number of candidate events as their test statistic, but instead the quantity $d(\mathbf{X}) = -2\ln\frac{\lambda(\mu_0|\mathbf{X})}{\sup\{\lambda(\mu_1|\mathbf{X})\}}$ (Feldman and Cousins 1998). This statistic uses the *likelihood function* $\lambda(\mu|\mathbf{X})$, a function that assumes different values for various values of $\mu \in \mathrm{M}$ for any particular value of $\mathbf{X}$, such that $\lambda(\mu|\mathbf{X}) \equiv Pr(\mathbf{X}; \mu)$, where $Pr(\mathbf{X}; \mu)$ is the probability distribution of $\mathbf{X}$ given a value $\mu \in \mathrm{M}$. The quantity $d(\mathbf{X})$ is thus defined, for data $\mathbf{X}$, in terms of the likelihood for the null hypothesis $\lambda(\mu_0|\mathbf{X})$ and the least upper bound (supremum) of the likelihoods of the alternative values of μ, $\sup\{\lambda(\mu_1|\mathbf{X})\}$. These quantities are single-valued for any given data $\mathbf{X}$, and the test statistic itself will take greater values to the extent that $\sup\{\lambda(\mu_1|\mathbf{X})\}$ exceeds $\lambda(\mu_0|\mathbf{X})$.

ATLAS provides a statistical characterization of the extent to which the number of Higgs candidate events in their data exceeds the expected contribution from background on the basis of the significance test just described.

In fact ATLAS provides multiple statistical characterizations of that excess. The reason for this is related to the fact that the distribution of the test statistic under the null hypothesis itself is not uniquely defined. The sensitivity of the experiment to the presence of decays of Higgs bosons depends in part on an unknown parameter: the mass of the Higgs boson m_H. Assuming that Higgs bosons do exist, the rate at which they are

produced is a decreasing function of m_H. This bears on the definition of the test statistic, in which the likelihood function for the alternative hypothesis H_1 appears in the denominator. Put differently, m_H is a *nuisance parameter* in the Higgs search, a parameter on which the sampling distribution for $d(\mathbf{X})$ depends, but that has an unknown value. Both CMS and ATLAS faced this difficulty, and dealt with it using somewhat different implementations of the same strategy, which is to begin by regarding the p value as a function of the parameter m_H. For each value of m_H there is a p value that is local to it (the *local* p value). The next step is then to report that function (see Fig. 1b).[4] ATLAS reports that the maximum significance (minimum local p value) of 6.0σ is achieved with the hypothesis of a SM Higgs boson with mass $m_H = 126.5$ GeV (taking systematic uncertainties into account lowers it to 5.9σ). However, the value of m_H, being unknown at the outset, is not in fact set in advance. A similarly significant result anywhere within the region of Higgs masses to which the experiment was sensitive would have yielded a similar claim of observation. The probability that the experiment would report an excess as great as that observed *for some value or other* of the Higgs mass, assuming the null hypothesis is true, is therefore greater than the minimum local p value. This discrepancy, elsewhere known by names such as the "multiple trials" or "multiple tests" effect, is known in HEP as the "Look Elsewhere Effect" (LEE).[5]

The size of that discrepancy depends on the range of values of m_H that one considers, and just what that range should be is not uniquely defined. Both ATLAS and CMS reported, along with their local p values, *global* p values, which report the probability of finding an excess anywhere within a range of possible values of m_H. Both groups, in order to emphasize that these ranges are "arbitrary or subjective" (Cousins 2014, p. 33), reported both "wide" and "narrow" ranges, based on different criteria. Cousins notes, "Some possibilities were the range of masses for which the SM Higgs boson [had] not previously been ruled out at high confidence; the range of masses for which the experiment is capable of observing the SM Higgs boson; or the range of masses for which sufficient data had been acquired to search for any new boson. The experiments made different choices" (ibid.). They certainly did. The range reported by ATLAS as "narrow" runs from 110 to 150 GeV (with a significance of 5.3σ), while the range reported by CMS as "wide" is 110–145 GeV (with a significance of 4.5σ). Reporting these global significance values serves as a kind of check on the sensitivity of the statistical significance to the LEE, and to fulfill that purpose the end points of the ranges used need not be uniformly determined.

The kind of sensitivity analysis conveyed by the reporting of multiple global p values addresses uncertainty about how precisely to model the experiment that has been performed. For any particular model, the p value is perfectly well-defined, but a number of different models are plausible. The task is to show that the p value does not depend strongly on just which of those plausible models is chosen (Staley 2002). Moreover, the reporting of multiple ranges (as well as showing how the local p value varies with m_H) serves an important communicative function, enabling the reader of

[4] As Cousins states, "for each mass [m_H] there is a p value for the departure from H_0, *as if that mass had been fixed in advance*" (Cousins 2014, p. 33, emphasis in original).

[5] For a discussion of the LEE in the Higgs search from a Bayesian standpoint, see Dawid (2015a, b).

the experimental report to consider for herself the sensitivity of the results to different ways of thinking about the experimental search. Insofar as different readers might have interests and beliefs that lead them to consider different ranges to be relevant, they might be interested in different "elsewheres."[6]

On the view here advanced, the p value plays an important *argumentative* role in establishing the existence of a new boson. What enables it to play this role is that it quantifies one dimension of a multi-dimensional evaluation of the evidence supporting that claim: It is important, for the purpose of cogently arguing for their claim, that ATLAS be able to establish that they have taken sufficient care to rule out, on a reasonable basis, the possibility of being misled by a stochastic fluctuation in the background. The calculation of a p value addresses that need. Other dimensions of ATLAS's assessment of the evidence that are essential to their experimental argument include (1) the distribution of the candidate events across different decay modes, (2) the comparison of the data with theoretical expectations for a Higgs boson with a mass in the range indicated by the data, (3) the ability to arrive at an estimated mass for the candidate decay events, and (4) the comparison of the estimated mass for different decay channels.

3 Warrant for significance testing

What makes the warrant for significance testing in HEP *pragmatic* can best be seen by contrast. One might suppose that the choice of a probabilistic framework for data analysis should be made on the basis of whether the framework yields verdicts regarding the credibility of hypotheses that align with our pre-theoretical views about when we have good grounds for—i.e., are epistemically justified in—believing a hypothesis. When the warrant for a choice of probabilistic framework has this character, we might call the warrant *epistemic*. The warrant here discussed differs in that its relationship to matters of belief is indirect.

My argument assumes that the ATLAS and CMS collaborations aim (collectively) not (only) to form beliefs about scientific subjects, but to contribute to the *production of scientific knowledge*. This assumption attributes to these groups a fundamental aim that is simultaneously epistemic (insofar as it concerns knowledge) and pragmatic (insofar as it concerns a productive activity undertaken collectively).

That productive activity involves, among other things, the practical tasks of discovery and argumentation. The warrant for the use of significance testing in HEP is pragmatic in the sense that significance testing is useful for the pursuit of these practical tasks. That discovery, in particular, should be thought of as practical requires some argument, which I now undertake.

The practical nature of these tasks might best be appreciated by viewing them in light of three questions, the answers to which bear on the manner in which experimental data will be analyzed: (1) What are the learning goals of the experiment? (2) What are the possible errors that must be confronted? and (3) What are the foreseeable practical

[6] I have adopted this felicitous expression from a comment by a referee.

consequences of those errors or their absence, including those that bear on further and related inquiries?

As these three questions hold the key to the pragmatic warranting of ATLAS's statistical practices, it is worth taking them in turn.

(1) What are the learning goals of the experiment? Obviously, ATLAS aimed to answer the question "Does the Higgs boson exist?" But they sought to do more than this. They sought to enable themselves to base their answer to that question on their own experimental data, and to do so in such a way that providing a positive answer would constitute a *discovery*. Discovery involves more than belief. As Dawid has noted (Dawid 2015a, b), many physicists, perhaps most, already thought themselves well warranted in assigning a high degree of belief to the Higgs hypothesis. ATLAS sought either to *discover* the Higgs boson (if it exists) or to rule it out (if it does not). Moreover, they sought to answer the Higgs question in a way that would enable them to make a persuasive *argument* to their intended audience in support of the answer at which they arrived and to support any discovery claim that might accompany that answer. (It deserves notice here that, especially in the context of a large collaboration like ATLAS, this argumentation aim is directed both externally and internally. The former is based on the need to present a cogent case for discovery to others, the latter is based on the need for establishing consensus within the group; see Rehg and Staley 2008.)

To make a discovery claim is not merely to report a high degree of belief in a hypothesis that asserts the existence of some new phenomenon, but to declare oneself in possession of data or evidence that provides significant new support for that hypothesis. Such new support cannot be based on an experimental test with results that are just what one would expect if the phenomenon in question did not exist.[7] To *discover a new phenomenon* requires a discrepancy between the experimental outcome in question and what one would expect in the absence of that phenomenon. An argument in support of such a discovery must accordingly establish the existence of such a discrepancy. Significance testing is thus useful for the purpose of discovery because it requires experimenters to (1) construct a model on the basis of which they may estimate what one should expect in the absence of the phenomenon in question, and (2) quantify the discrepancy (in a statistical sense) between that expectation and the actual experimental outcome, while also (3) confronting the question of the liability of their testing procedure to generate erroneous conclusions from that discrepancy. These same features make it useful for argumentative purposes because the explicit consideration of these three factors makes them available for deployment in an argument that seeks to establish that a discovery has been made.

(2) What are the possible errors that must be confronted? The third of the three features of significance testing just mentioned—the liability of a test to error—is clearly addressed in significance testing at least in the sense of quantifying the probability of a Type I error of rejecting the null hypothesis, supposing it is true. Although the probability of a Type II error is not explicitly calculated in the significance test procedure, the physicists searching for the Higgs were clearly motivated to avoid the problem of

[7] The principle invoked here is similar to what Mayo and Spanos call the "weak severity principle" (Mayo and Spanos 2009, p. 21).

having a high probability of failing to reject the null hypothesis, assuming it to be false. The choice of the likelihood ratio as a test statistic itself draws its justification from the desirable Type II error probabilities of tests based on the likelihood ratio. Another way in which the probability of a Type II error could be made too high would be by setting the discovery threshold for significance too high (as discussed in the next section).

Thinking of the possible errors simply as accepting the existence of the Higgs when only background processes are present and failing to accept the Higgs hypothesis when the Higgs is present provides us with only the coarsest-grained description of the landscape of errors surveyed by the ATLAS physicists. The complexity of that landscape reflects the complexity of the analytic procedures ATLAS brought to bear on their data. For the purpose of simply assessing the warrant for ATLAS's reliance on significance testing it will have to suffice that these two primary ways of going wrong are the ultimate source of concern, and all of the more fine-grained possibilities of error become relevant precisely because of their potential to lead to one or the other of these two main errors.

At this point, however, it is important to address a concern about frequentist methods such as significance testing.[8] Phrases like "the probability of committing an error" are ambiguous between two distinct quantities, only one of which is addressed, or even considered legitimate, in frequentist statistics. The first quantity, for the outcome of a significance test, is $Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0); H_0)$, which is the probability of obtaining a value for the test statistic that is at least as great as that obtained from the observed data, assuming the null hypothesis is true. This, of course, is simply the p value. The second quantity that we might regard as "the probability of committing an error" in a test of a null hypothesis is $Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0) \wedge H_0)$. For frequentists the latter probability is illegitimate because it requires the determination of a probability for the null hypothesis itself. Thus, a warrant for relying on significance testing cannot rest simply on demonstrating that it addresses a concern with limiting the probability of erroneously rejecting the null, since it only relates to one of the two ways of conceiving that error.

In response to this point, recall that I have emphasized the use of the significance test as playing an important argumentative role in supporting ATLAS's claim to have discovered a new Higgs-like boson. That role, specifically, is to establish the existence of a sufficient statistical discrepancy between the background hypothesis and the excess of candidate Higgs decays in the data. Although the LEE introduces what some might regard as an ambiguity in the definition of the p value for the Higgs results, the sensitivity analysis reported via the calculation of global p values lends credibility to the effectiveness of the p value as a measure of statistical discrepancy by showing it to depend only very weakly on the precise model of the experiment performed. The p value remains a useful device for its argumentative purpose.

The probability $Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0) \wedge H_0)$ would not be suitable for this same purpose, for two reasons.

First, its calculation would require determining a value for the prior probability $Pr(H_0)$. The difficulty here is not that there is no way to do this, but that there are too many ways to do it, from the 'reference priors' of objective Bayesians to elicitation

[8] I am grateful to an anonymous referee for pressing this issue.

procedures aimed at discovering the personal probabilities of experts. Of course, one could simply give $Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0) \wedge H_0)$ as a *function* of $Pr(H_0)$ and let the reader choose her own prior probability. Indeed, such a report could be useful, but were experimentalists to limit themselves to reporting *only* this, they would be abandoning one of their own central aims, radically revising the task of reporting the outcome of an experiment. Bayesians commonly criticize p values on the grounds that "what we really want to know" is not the probability of getting such-and-such a result assuming that the null hypothesis is true but the probability, in light of the data, that the null (or alternative) hypothesis is actually false (or true). But even if it is true that this is "what we really want to know," it does not follow that we should abandon significance testing for Bayesian statistics. Bayesian analyses cannot deliver this quantity either. Instead, they can only tell us what posterior probability we *would* arrive at, were we to begin with any particular prior probability.

Second, even were it possible to determine a unique value for $Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0) \wedge H_0)$ (or, as a Bayesian would prefer, the function $Pr(\mathbf{X}|H_0)Pr(H_0)$, evaluated at $\mathbf{X} = \mathbf{x_0}$), it would still not serve the argumentative function played by p values. Because it is equal to the product of the p value (interpreted now as a conditional probability) and the prior probability of H_0, the fact that this number is low would not by itself indicate a statistical discrepancy between H_0 and the results of the test, since such a low number could simply reflect a low value for the prior probability. The part of the calculation relevant for argumentative purposes would remain the p value.

ATLAS's choice of a methodology of significance testing is warranted in light of their aims and in light of the kinds of errors they sought to avoid. They sought to be able to make a clear and compelling case, for a somewhat diverse audience, either for or against the existence of the Higgs, while limiting the probability of doing so erroneously. Significance testing alone is not sufficient for this aim, but it can contribute to the pursuit of it, by providing evidence regarding the compatibility of the data with the background-only hypothesis.

To clarify the contrast between the approach taken here and an epistemic approach, consider an example of the latter. Richard Dawid's recent discussions of the Higgs discovery (Dawid 2015a, b) are noteworthy for their sensitivity to many of the issues that have been raised in this essay. Dawid notes the considerable confidence that many physicists had in the Higgs hypothesis prior to the July 2012 announcements, and raises the question whether this confidence should make a difference to the way in which the experimental data are analyzed [particularly regarding the Look Elsewhere Effect (LEE)] and the standards of discovery that should be applied.

Dawid distinguishes two positions. The *experimentalist's position* maintains that prior confidence in the Higgs hypothesis does not warrant treating the data differently than one would in other experimental searches for new phenomena, and that the same standards of discovery should be applied. The *theoretician's position* claims that prior confidence in the Higgs hypothesis is well warranted and thus can and should be taken into account in the treatment of experimental data. This difference of opinion Dawid traces to a difference of priorities: the experimentalist's position "has the priority to defend the purity of the process of data analysis by keeping it free of theoretical reasoning that is about to be tested by those very data" while the theoretician's position

"has the priority to be frank about our actual beliefs with respect to the hypothesis in question" (Dawid 2015b, p. 83).

In his (2015b), Dawid employs the theorist's position to defend a partially Bayesian approach to the Higgs data. In effect, he argues that Bayesian considerations of the plausibility of the Higgs hypothesis in light of both theoretical considerations and previous experimental results warrant a limitation of the LEE that obviates the need for a strict adherence to a 5σ standard of discovery. Dawid's proposal retains the use of significance testing, insofar as the criterion of discovery still refers to the local p value as the relevant quantity (as opposed to the reporting of a posterior probability as called for in a fully Bayesian approach). The standard applied to that quantity, however, is subject to modification based on Bayesian considerations of the prior probabilities of the relevant hypotheses.

The syncretism of Dawid's approach resembles that advocated here in that it avoids allowing rigid adherence to a single probabilistic framework to dominate decisions about how best to analyze data, but his syncretism is not grounded in pragmatism in the same sense as here proposed. Dawid instead advocates a resolutely Bayesian epistemic framework: "Any coherent characterization of the transition from a phase of non-empirical theory confirmation to the discovery of corresponding particles must be based on a Bayesian overall perspective as well, which speaks in favour of an epistemically Bayesian embedding of frequentist data analysis" (Dawid 2015a, p. 17).

Dawid's allowance of frequentist methods arises from a deference to physicists' preference for objectivity: completely replacing frequentist statistical analysis of data with a completely Bayesian approach would "permit that a rigid quantitative statistical analysis where the numerical input is well determined by the empirical data gets adulterated by probability assessments that are vague and subjective. It would put guessing priors on the same footing as rigid and quantitative experimental testing," whereas retaining a frequentist approach to data analysis "avoids messing with the objective character of statistical data analysis itself" (Dawid 2015b, p. 92).

Dawid is certainly correct to point to the value of objectivity as a consideration that physicists cite in favor of frequentist statistics and against full-fledged Bayesianism, but leaves unaddressed the question of when and how objectivity matters to the decisions scientists must make about how to analyze their data.

According to the view advanced here, the value of significance testing rests not merely on a preference for objectivity, but on its suitability for certain practical tasks of experimental HEP. It seems likely that the value of objectivity in significance calculations itself rests at least in part on the *argumentative effectiveness* of ways of characterizing the data that depend as little as possible on propositions that are subject to differences of opinion.

4 The 5σ standard

(3) What are the foreseeable practical consequences of those errors or their absence, including those that bear on further and related inquiries? Knowing what would constitute an error is not the same as knowing what will happen once an error is committed. This question receives no explicit treatment in ATLAS's published Higgs

results. This, however, does not mean that consideration of it played no identifiable role in their deliberations over the statistical assessment of their data. On the contrary, the consequences of erroneously announcing a discovery of the Higgs played an important role in their reliance on what many regarded as an extremely strict standard of significance: the "5σ" rule previously mentioned. As this may constitute the clearest instance of pragmatic thinking in this episode, this point deserves its own discussion.

Although the requirement that discovery claims in HEP be premised on statistical excesses that are significant at a level of 5σ has assumed the status of tradition within the HEP community,[9] it has no official institutional codification and physicists will deny that its normative force is absolute. According to Joe Incandela, who was spokesperson for CMS at the time of the July 2012 announcements, "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, p. 30). Meanwhile, some physicists have called for reform of the 5σ standard. Louis Lyons, for example, has called for a "more nuanced criterion" that would be more or less demanding for a variety of possible future discoveries, based on four criteria: the presence of an LEE, the magnitude of systematic uncertainties, the impact of the discovery, and the "degree of surprise" (also called the "subconscious Bayes' factor") (Lyons 2013).

Lyons' criteria cohere well with responses that Tony O'Hagan received from physicists to his query regarding the rationale for the 5σ criterion, mentioned in Sect. 1. Acknowledging the statistical (and pragmatic) inappropriateness of an ironclad significance threshold for discovery claims, these responses (apart from a minority of Bayesian physicists calling for the abandonment of significance testing altogether) indicated an acceptance of 5σ as an appropriate standard for the Higgs search itself. (In Lyons' enumeration of varying significance standards from 3 to >8 standard deviations for fourteen different HEP searches, the standard for the Higgs search remains at 5σ.) Prominent among the considerations cited are the LEE[10] and systematic uncertainty, or more generically, to quote O'Hagan's summary, the fact that "so much can go wrong that it makes sense to guard against false positives caused by errors in underlying assumptions, pre-processing, experimental controls, etc." (O'Hagan 2012, p. 5).

The problems of the LEE and systematic uncertainties constitute obstacles toward taking the calculated local significance seriously as an accurate measure of what it purports to be: the probability of observing an excess as great as or greater than that

[9] Allan Franklin has documented the emergence of the 5σ standard in HEP (Franklin 2013). According to Franklin's narrative, the standard has only assumed the weight that it does carry rather recently, around the time of the discovery of the top quark, for which an initial paper by CDF in 1994 claimed only "evidence," with a significance corresponding to 2.8σ for a Gaussian distribution (Abe et al. 1994). Later papers by CDF and D0 claimed the top's "observation" on the basis of 5.0σ and 4.6σ, respectively (Abe et al. 1995; Abachi et al. 1995; Staley 2004).

[10] As noted above, however, Dawid (2015a) uses Bayesian considerations to argue that the LEE is not as significant a problem for the Higgs search as others have suggested, and therefore application of the stringent 5σ standard is unjustified.

observed, assuming that only background processes are present. They leave unaddressed the further questions of why the standard for discovery should be a very demanding one in the first place (why is 3σ not good enough?) and why it should not be even more demanding (why would 8σ not be even better?). Dawid argues that Lyons' "subconscious Bayes' factor" is relevant to these questions (Dawid 2015a, b). Here I consider how the criterion of *impact* bears on them. The impact of the outcome of the Higgs search can be illuminated by addressing question (3): What are the foreseeable practical consequences of the possible errors or their absence? Although answers to this question do not determine univocally a precise standard that must be applied, they will illuminate the reasons that shaped the terrain in which the decision was made.[11]

The pragmatic perspective requires us to acknowledge that the outcome of an inference is not only an event in an abstract realm of ideas, but is a decision with practical consequences. As C. West Churchman notes,

> 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. (Churchman 1948, p. 259)

It is commonly held that inference and decision are distinct kinds of problems calling for distinct analytic frameworks.[12] The position here advocated does not dispute that there is an important distinction between treating a problem as a matter of inference and as a matter of decision. Nor should one neglect the value of analyzing the data as though one were faced with a strictly inferential question. Rather, the point emphasized by Churchman is that, as regards the practice of engaging in scientific research, a purely inferential perspective is *incomplete*. The analysis of data makes a difference to the state of scientific knowledge only via the decisions of researchers regarding what to report on the basis of that analysis, how to report it, and when to report it. Such decisions cannot be made without some consideration of their potential consequences, for good or ill. It may be that in many cases the consequences are so transparent or so inconsequential that no explicit discussion of them is called for. For a coherent philosophical understanding of the scientific process, acknowledgement of the role and import of such decisions is essential nonetheless.

In other words, even if scientists individually engage in what Isaac Levi calls "attempts to seek the truth and nothing but the truth" (Levi 1962), the conduct of experimental inquiry is not only a matter of the pursuit of true beliefs and the avoidance of false ones. Because science is an essentially social undertaking, it necessarily

[11] Those familiar with the "argument from inductive risk" that seeks to establish a role for value judgments in core tasks of scientific reasoning (Churchman 1948; Douglas 2009; Rudner 1953) will note its resemblance to the point being here pursued. See Staley (2016) for further discussion of inductive risk in the context of the Higgs search.

[12] Exactly which framework is appropriate for which problem remains a matter of dispute. *Statistical Methods in Experimental Physics*, a widely used text, emphasizes (in both of its two editions) frequentist techniques for the analysis of data, and introduces Bayesian statistics as an approach to decision problems (alongside frequentist methods) (Eadie et al. 1971; James 2006).

involves decisions about how to carry out argumentative and communicative tasks, and the statistical analysis of data is an element of such tasks. It may make perfect sense to attribute to ATLAS and CMS the aims of accepting true propositions and rejecting false ones, but the successful pursuit of these aims is not sufficient for science, which requires also that significant truths be communicated clearly and supported with cogent argumentation. Communication and argumentation are *practical* tasks, albeit ones with potentially significant epistemic import (Staley 2016).

The decisions implicated in the discovery of a new boson in July 2012 took place at various levels and were distributed across various actors. In addition to all of the intra-group decisions regarding analysis, discussion, and argumentation, the groups as a whole had to reach a decision to go forward with the submission of papers declaring that a new boson had been "observed." Beyond those decisions lay the decision of the lab director, Rolf Heuer, who made the actual announcement and whose reasoning drew upon the fact that *both* groups had independently accumulated results with a 5σ significance.

We can place the consequences of these decisions into two categories: those that pertain directly to the logical argumentation of future physics inquiries, and those that pertain indirectly to the aims of ATLAS, CMS, and the HEP community more broadly.

Regarding the first category, accepting the existence of a new boson amounts to a commitment to adopt statements entailing the existence of such a particle as premises in the pursuit of further inquiries. This commitment has its most obvious salience for the continued work of ATLAS and CMS themselves, as their analytic tasks turn from the aim of producing exclusion plots towards the aim of measuring the properties of the newly discovered particle and probing further implications of the Higgs hypothesis to fix more securely the theoretical interpretation of their finding. For other physicists working on SM and Beyond-SM problems, the announcements by ATLAS and CMS change the logical terrain. Although each investigator must decide (as an individual or as a member of a working group) whether the evidence offered by the two CERN groups suffices to warrant agreement with their discovery claims, 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 within ATLAS and CMS as well as beyond, 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 previously had to wait offstage.

The second category of consequences must be regarded as somewhat more speculative, but various statements of physicists involved in the Higgs search suggest some relevant considerations. 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. 2012b, p. 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 themselves were not merely attended by submitting papers for publication, but by a kind of scientific showmanship including a presentation to the press 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.

Although one cannot be certain of the consequences of such an error, it is not unreasonable to imagine them including even a political dimension with negative consequences for the funding of HEP.

One respondent to O'Hagan's query communicates vividly the personal nature of such considerations: "In fact, we do have high standards because in our view we are trying to arrive at 'true' statements about the world in the pragmatic sense that these statements yield predictions that turn out to be correct. Given that the search for the Higgs took some 45 years, tens of thousands of scientists and engineers, billions of dollars, not to mention numerous divorces, huge amounts of sleep deprivation, tens of thousands of bad airline meals, etc., etc., we want to be sure as is humanly possible that this is real" (O'Hagan 2012, p. 5).

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).

One might at this point object that the warrant I have been discussing is not really pragmatic, but ultimately, or at least primarily, epistemic. The aims that are concerned directly with inquiry are clearly concerned with the pursuit of knowledge. Even concerns about the consequences of error for the status and funding of HEP or for the public perception of science are really concerned with sustaining the scientific pursuit of knowledge.

This objection, however, misunderstands the pragmatic perspective on inquiry here proposed. The broad sense of 'practical' that I have invoked should be understood to include those actions that are part of the scientific pursuit of knowledge. This pragmatic perspective on inquiry highlights the fact that inquiry is not directed at truth alone, but at the ability to communicate and argue on behalf of propositions that answer questions that are part of an investigation undertaken deliberately, systematically, and collectively. The epistemic is shot through with the pragmatic. Scientists use procedures that are warranted not because they guarantee the epistemic warrant of the propositions thus advanced (which always remains an empirical question to be decided on a case-by-case basis), but because of their strategic value in the pursuit of more specific aims (in this I am in agreement with the view recently advanced by Achinstein 2013). In the present example, the aim I have been discussing is the determination of a statistical discrepancy between the data and the null hypothesis, for the purpose of providing an important premise in arguing for a discovery claim in HEP, and the strategically valuable rule may be formulated as 'report a p value and accept the discovery claim *only if* the significance exceeds the 5σ level.'

Taking the pragmatic perspective allows us to see that considerations regarding the consequences of an inference are not extraneous to the scientific process, but rather help to clarify it. A clear articulation of the meaning of an inference will bring to

light its practical dimension, thus helping us to understand the evidential standards that have been brought to bear on it—standards that might otherwise seem arbitrary or mysterious.

5 An objection

My argument for the pragmatic warrant of significance testing in HEP has emphasized the importance, for argumentation purposes, of establishing a statistical discrepancy between the background hypothesis and the observed excess of Higgs candidates, and the suitability of calculating a p value for doing so. This justification would clearly be circular, were the perceived importance of this argumentative task itself an artifact of the adoption of the methodology of significance testing.

I contend in response that we have good reason to believe that the importance, in arguing for a discovery claim in HEP, of establishing a statistical discrepancy between the null hypothesis and the observed data is logically *prior to* the choice of statistical methodology.

I have already cited a principle that supports this claim of logical priority: Discovery requires data or evidence that provides significant new support for the existence of a new phenomenon, and that requires, in turn, that such data yield a discrepancy with what one would expect if the phenomenon in question did not exist. A compelling argument to support the experimental discovery of a new phenomenon should establish the existence of such a discrepancy.

Adherence to this guiding principle is reflected in the history of particle physics experimentation, prior to the widespread adoption of a significance testing methodology. Consider, for example, Carl Anderson's 1933 discovery of the positron. Anderson begins his paper announcing this discovery with discussion of a single photograph from his cloud chamber. In the center of this photograph (see Fig. 3) one can see the image of the lead plate that cuts across the center of the cloud chamber, with a track that seems to pass through the lead plate. The curvature of the track is attributable to the action of a magnetic field on a charged particle. This image, he argues, can only be interpreted as a "particle carrying a positive charge but having a mass of the same order of magnitude as that normally possessed by a free negative electron" (Anderson 1933, p. 491). This interpretation, he argues, is "inevitable." Anderson's argumentation is heterogeneous: some arguments target alternative explanations as being untenable or implausible at the outset; other arguments aim to establish, in an informal way, a statistical incompatibility between an alternative explanation and particular features of the photograph. He compares those features with expectations derived from two possible alternative interpretations based on the only charged particles known at the time: the negatively charged electron and the proton:

The change of curvature due to loss of energy in the lead plate indicates that the particle was traveling from the bottom of the cloud chamber to the top. Based on the known orientation of the magnetic field, Anderson concludes that the particle has a positive charge. Perhaps it is a proton? Anderson determines the energy such a proton would have from the curvature of the track. The total range in air of a proton with that energy is an order of magnitude smaller than measured length of the track in the

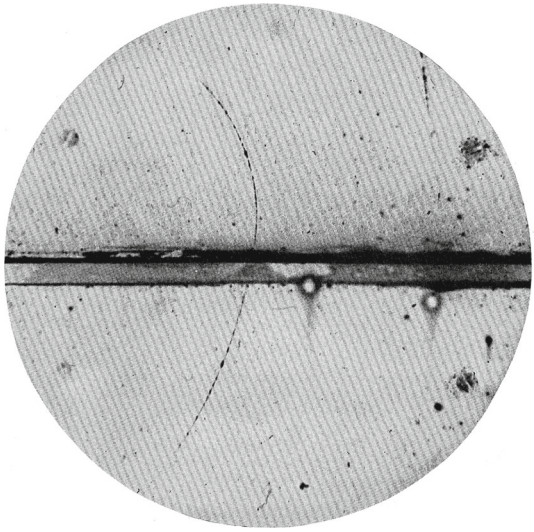

Fig. 3 A cloud chamber photograph of a track left by a positron. The particle enters from below and passes through a 6 mm lead plate across the center. The subsequent loss of momentum results in the greater curvature of the track in the upper region. (Anderson 1933, p. 492)

chamber. The observed feature is, thus, incompatible with the expectation from the proton hypothesis.

Anderson then turns to the consideration of a negative electron interpretation of the track. He considers two scenarios: (1) two negative electrons just happened independently to line up so as to produce the appearance of a single particle passing through the lead plate; (2) a single negative electron, traveling from the top of the chamber to the bottom, made the track. Anderson rules out scenario (2) as implausible, since it would require the electron in question to *gain* 40 million electron volts from passing through 6 mm of lead.

Anderson's only explicit (though still informal) reference to probability comes in his argument against the negative electron scenario (1): "This assumption was dismissed on a probability basis, since a sharp track of this order of curvature under the experimental conditions prevailing occurred in the chamber only once in some 500 exposures, and since there was practically no chance at all that two such tracks should line up in this way" (ibid.). (Anderson had a total of 1300 photographs.) Here Anderson argues that under the assumption that only negatively charged electrons are involved in producing the track, the probability of an image such as that in Fig. 3 is very low. He does not actually calculate that probability. From the "once in 500" number that he does cite, we might estimate the probability of finding two such tracks in the same photograph as 4×10^{-6}. Adding the requirement that the two tracks line up exactly so as to produce the appearance of a single particle passing through the lead plate presumably reduces the probability to a considerably smaller value, thus justifying Anderson's "practically no chance at all" claim.[13]

[13] Anderson also discusses the possibility that an undetected photon struck a nucleus in the lead plate, knocking out two particles in opposite directions. This, however, does not really count as an alternative to

The argument just summarized is central to Anderson's claim to have discovered the positron. He clearly judged it important to establish the incompatibility of the data he had collected with the denial of the hypothesis for which he claimed support. Just as clearly, this judgment was not motivated to some prior commitment to the methodology of significance testing. Fisher's *Statistical Methods for Research Workers* had only been published 8 years previously (Fisher 1925), and (as Allan Franklin documents in his 2013) physicists would not take up the systematic use of significance calculations until some time later.

Of course, Anderson's paper is only one example, and does not suffice on its own to establish that particle physics argumentation in general abides by the principle that supporting discovery claims requires establishing the statistical incompatibility of data with the background hypothesis. A more thorough historical argument to this effect would require a more lengthy discussion than can here be afforded. Nonetheless, I contend that the argumentation in Anderson's paper establishes the plausibility of my historical claim. Moreover, the papers discussed in Franklin (2013) provide an excellent resource for finding further supporting evidence.

6 Concluding remarks

This paper has argued that pragmatism helps us to understand how the statistical methodology of HEP is warranted. Naturally, another philosophical framework that *could* warrant the use of such frequentist statistical methods would be statistical frequentism itself. Indeed, one could regard frequentism and pragmatism as compatible. Neyman and Pearson themselves sometimes seem to articulate ideas that at least seem compatible with pragmatism, and the influential early pragmatist C. S. Peirce himself articulated a resolutely frequentist anticipation of the Neyman–Pearson approach (Peirce 1883).

However, if we understand frequentism either as the (strong) position that the only probability statements that are meaningful are those involving probabilities as relative frequencies or as the (weaker) position that only probabilities understood as relative frequencies are useful in the statistical analysis of data, then we have to regard as problematic another aspect of HEP statistical practice, which is its incorporation of Bayesian techniques in, typically, supporting roles in the analysis of data. An example of this concerns a commonly used technique for incorporating systematic uncertainties, such as when an estimate of a physical quantity requires assigning a value to another, auxiliary, physical quantity that is imperfectly known. One solution to this problem involves assigning a probability distribution to the auxiliary quantity, in effect dispersing the estimated quantity across a broader range of values than one would obtain from assigning a fixed point value to the auxiliary quantity (Cousins and Highland 1992). The probability distribution assigned to the auxiliary quantity cannot be given a frequentist interpretation (but see Willing 2013).

Footnote 13 continued
the claim of a positively charged electron, since (from considerations of curvature and direction) one of the particles knocked out of the lead plate would have to be just such a positively charged, low-mass particle.

In cases such as these, pragmatism seems to trump frequentism. The commitment to frequentist statistics apparently does not rest on a belief that only statements about frequency probabilities are meaningful or useful. Although a careful discussion of such methods for dealing with systematic uncertainties is a subject for another paper, pragmatism points us toward the kinds of considerations that would be relevant to understanding such an apparent statistical eclecticism. For example: What is the epistemological problem to be solved? What are the argumentation requirements for a satisfactory solution to this kind of problem? How can solutions to this kind of problem be related to the results of work undertaken on connected scientific problems?

The argument of this paper has focused on the particular case of the use of p values in the argument for the discovery of a new Higgs-like boson based on the Higgs search results at ATLAS and CMS. In spite of the many criticisms of p values, the LHC physicists' use of them was warranted because they employed significance testing for the specific purpose of providing evidence relevant to the multi-dimensional assessment of the hypothesis that their excess of Higgs candidates was due to a stochastic fluctuation of non-Higgs background processes. Their use of significance testing was tailored to specific inferential and argumentative aims, in light of explicit consideration of the possible errors that could be made in drawing an inference, and with at least implicit attention to the consequences of such errors, both for immediate matters of related scientific inquiries and for broader matters related to the place of HEP and science in society.

I have argued here that considerations of the consequences of possible errors of inference played an important (though not exclusive) role in the determination of standards of evidence for purposes of announcing a discovery based on the Higgs search results at LHC. Discussions of the practical consequences of accepting a hypothesis are part of the pragmatic clarification of an inference. Yet current norms governing scientific communication tend to force such discussions into informal, background contexts, so that the resulting decisions appear to the public as they were reported in the press following the Higgs announcement of July 2012: as a "gold standard" or as a "strict notion of scientific certainty" the status of which is simply to be taken for granted.[14] Although I would not propose that every positive scientific claim must be accompanied by a detailed discussion of the deliberations that guided the choice of evidential standard that was applied to that claim, I do think that a more complete execution of the program of pragmatic clarification should include a more systematic expectation that scientists in fields such as HEP should address explicitly and thoroughly the considerations—including those regarding potential consequences of errors—that guide such decisions.

Acknowledgements I would like to thank Richard Dawid for his editorial patience and encouragement. This paper grew out of a presentation at a conference organized by Michael Stöltzner, and I would like to

[14] I am not denying that CMS and ATLAS gave detailed presentations; on the contrary, they accompanied a careful evidential argument with very thorough presentations of the analyses they applied to the data. What I am claiming is that conventions of scientific communication tend to suppress explicit discussion of those considerations of the consequences of accepting a claim that play a role in determining the evidential standard of acceptance.

thank Michael and other participants, including Richard Dawid, Hugo Beauchemin, Robert Cousins, and Koray Karaca for their insights into these issues. Conversations with Deborah Mayo and Allan Franklin were also helpful. I am especially grateful to three anonymous referees whose generous comments on earlier drafts were very helpful in improving this paper.

References

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., et al. (2012a). Combined search for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Physical Review D*, *86*, 032003. doi:10.1103/PhysRevD.86.032003.

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Abdel Khalek, S., Abdelalim, A. A., et al. (2012b). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, *716*, 1–29. doi:10.1016/j.physletb.2012.08.020.

Abachi, S., Abbott, B., Abolins, M., Acharya, B. S., Adam, I., Adams, D. L., et al. (1995). Observation of the top quark. *Physical Review Letters*, *74*, 2632–2637. doi:10.1103/PhysRevLett.74.2632.

Abe, F., Akimoto, H., Akopian, A., Albrow, M. G., Amendolia, S. R., Amidei, D., et al. (1995). Observation of top quark production in $\bar{p}p$ collisions with the collider detector at Fermilab. *Physical Review Letters*, *74*, 2626–2631. doi:10.1103/PhysRevLett.74.2626.

Abe, F., Albrow, M. G., Amendolia, S. R., Amidei, D., Antos, J., Anway- Wiese, C., et al. (1994). Evidence for top quark production in $\bar{p}p$ collisions at $\sqrt{s} = 1.8$ TeV. *Physical Review D*, *50*, 2966–3026. doi:10.1103/PhysRevD.50.2966.

Achinstein, P. (2013). *Evidence and method: Scientific strategies of Isaac Newton and James Clerk Maxwell*. New York: Oxford University Press.

Anderson, C. (1933). The positive electron. *Physical Review*, *43*, 491–494.

ATLAS. (2012). Latest results from ATLAS Higgs search (Press Release).

Chatrchyan, S., Khachatryan, V., Sirunyan, A., Tumasyan, A., Adam, W., Bergauer, T., et al. (2012a). Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV. *Physics Letters B*, *710*(1), 26–48.

Chatrchyan, S., Khachatryan, V., Sirunyan, A., Tumasyan, A., Adam, W., Bergauer, T., et al. (2012b). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, *716*, 30–61. doi:10.1016/j.physletb.2012.08.021.

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

CMS. (2012). Observation of a new particle with a mass of 125 GeV (Press Release).

Cousins, R. D. (2014). The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, 1–38. 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.

Cox, D. R. (1970). *Analysis of binary data*. London: Methuen.

Dawid, R. (2015a). Bayesian perspectives on the discovery of the higgs particle. *Synthese*, 1–18. 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. (2009). *Science, policy, and the value-free ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Eadie, W. T., Dryard, D., James, F. E., Roos, M., & Sadoulet, B. (1971). *Statistical methods in experimental physics*. Amsterdam: North Holland.

Feldman, G. J., & Cousins, R. D. (1998). Unified approach to the classical statistical analysis of small signals. *Physical Review D*, *57*, 3873–3889. doi:10.1103/PhysRevD.57.3873.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

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

Harlander, R. V., & Kilgore, W. B. (2002). Next-to-next-to-leading order Higgs production at hadron colliders. *Physical Review Letters*, *88*, 201801. doi:10.1103/PhysRevLett.88.201801.

James, F. (2006). *Statistical methods in experimental physics* (2nd ed.). Singapore: World Scientific.

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

Lyons, L. (2013). Discovering the significance of 5σ. arXiv:1310.1284.

Massimi, M., & Bhimji, W. (2015). Computer simulations and experiments: The case of the Higgs boson. *Studies in History and Philosophy of Modern Physics*, *51*, 71–81.

Mayo, D. G., & Spanos, A. (Eds.). (2009). *Error and inference: Recent exchanges on experimental reasoning, reliability, objectivity, and rationality*. New York: Cambridge University Press.
Morrison, M. (2015). *Reconstructing reality: Models, mathematics, and simulations*. New York: Oxford University Press.
O'Hagan, T. (2012). Higgs boson digest and discussion. Retrieved March 17, 2014, from http://bayesian.org/forums/news/3830.
Overbye, D. (2012, July 4). Physicists find elusive particle seen as key to universe. *New York Times*.
Peirce, C. S. (1883). A theory of probable inference. In C. S. Peirce (Ed.), *Studies in logic: By members of the Johns Hopkins University* (pp. 126–181). Boston: Little, Brown, and Company.
Rehg, W., & Staley, K. W. (2008). The CDF collaboration and argumentation theory: The role of process in objective knowledge. *Perspectives on Science, 16*, 1–25.
Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, *20*(1), 1–6.
Sprenger, J. (2016). Bayesianism vs frequentism in statistical inference. In A. Hajek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy*. Oxford: Oxford University Press.
Staley, K. W. (2002). What experiment did we just do? Counterfactual error statistics and uncertainties about the reference class. *Philosophy of Science*, *69*(2), 279–299.
Staley, K. W. (2004). *The evidence for the top quark: Objectivity and bias in collaborative experimentation*. New York: Cambridge University Press.
Staley, K. W. (2016). Decisions, decisions: Inductive risk and the higgs boson. In K. C. Elliott & T. Richards (Eds.), *Exploring inductive risk*. New York: Oxford University Press.
Venzon, D. J., & Moolgavkar, S. H. (1988). A method for computing profile-likelihood-based confidence intervals. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *37*(1), 87–94.
Wickham, C., & Evans, R. (2012, July 4). "It's a boson": Higgs quest bears new particle. *Reuters*.
Willing, R. (2013). *Measurement uncertainty and probability*. New York: Cambridge University Press.