Securing the Empirical Value of Measurement Results

Kent W. Staley

Abstract

Reports of quantitative experimental results often distinguish between the statistical uncertainty and the systematic uncertainty that characterize measurement outcomes. This paper discusses the practice of estimating systematic uncertainty in High Energy Physics (HEP). The estimation of systematic uncertainty in HEP should be understood as a minimal form of quantitative robustness analysis. The secure evidence framework is used to explain the epistemic significance of robustness analysis. However, the empirical value of a measurement result depends crucially not only on the resulting systematic uncertainty estimate, but on the learning aims for which that result will be used. Important conceptual and practical questions regarding systematic uncertainty assessment call for further investigation.

1 Introduction

The responsible reporting of measurement results requires assessing the quality of the measurement as well as its outcome. But the quality of a measurement is not one-dimensional. Standard practice in particle physics requires that all reports of measurement or estimation results must include quantitative estimates of both the statistical error or uncertainty and the systematic error or uncertainty. (This terminological ambivalence between error and uncertainty is addressed below. In the meantime, I will use 'uncertainty' to avoid awkwardness.) Such assessments of uncertainty are essential for the usefulness of measurement, for without them one cannot determine the consistency of two results from different experiments, two different results from the same experiment, or a single experimental result with a given theory (Beauchemin [2015]). Although a common practice of experimental particle physicists (and scientists in many other disciplines as well), the reporting of estimates of systematic uncertainty has received little attention from philosophers of science, apart from a few discussions noted in what follows.

Such neglect is unfortunate, for discussions of systematic uncertainty open a remarkable window into experimental reasoning. Whereas statistical uncertainty is simply reported, systematic uncertainty is also discussed. Even the most cursory presentation will at least note the main sources of systematic uncertainty, while more careful reports (such as the example discussed in the appendix) detail the ways in which systematic uncertainties arise as well as the methods by which they are assessed. Such discussions require forthright consideration by experimenters of the body of knowledge that they bring to bear on their investigation, the ways in which that knowledge relates to the conclusions they present, and the limitations on that knowledge. This process is epistemologically crucial to the establishment of experimental knowledge.

Moreover, philosophical insight regarding the estimation of systematic uncertainty would be highly valuable. Presently, there is no clear consensus across scientific disciplines regarding the basis or meaning of the distinction between statistical and systematic uncertainty, despite some concerted efforts mentioned below. Scientists likewise debate the proper statistical framework in which systematic uncertainty should be evaluated, a debate with important philosophical aspects.

It is the contention of this paper that some progress may come from regarding the estimation of systematic uncertainty as an instance of robustness analysis applied to a model of a single experiment or measurement. More precisely, the determination of systematic uncertainty bounds on a measurement result consists of a weakening of the conclusion of a model-based argument under the guidance of a robustness analysis of its premises within the bounds of what is epistemically possible.

Some measurement results have more empirical value than others, in the sense that they are more useful for promoting the learning aims of investigators, whatever those might be. Evaluation of systematic uncertainty is crucial to determining the empirical value of a measurement result, because it allows experimentalists to establish the sensitivity of the measurement result to phenomena of scientific interest.[1] At the same time, systematic uncertainty estimation promotes the security of the evidence that the data provide in support of the measurement result (in the sense of Staley [2004, 2012]), which is necessary for the cogency of the argument supporting the claim expressing the measurement result. However, I will argue, these two achievements are in tension with one another: as one weakens the conclusion to enhance the security of the evidence, one diminishes the sensitivity of the measurement result itself to the phenomena of interest. Just how much empirical value a measurement result maintains, therefore, depends not only on the extent to which the

[1]Measurement results may have empirical value in many ways. A very insensitive measurement result might, for example, reveal flaws in a novel measurement procedure. The sensitivity of a measurement result to the physics phenomena being investigated, however, addresses a central role for measurement in HEP: testing amongst competing substantive hypotheses relevant to the evaluation and development of theory.

sensitivity of the measurement has been weakened by systematic uncertainty bounds, but also on the use to which it will be put.

This account builds on two significant recent contributions to the philosophical study of measurement and measurement quality: Eran Tal's model-based account of measurement, according to which the evaluation of measurement accuracy is the outcome of a comparison amongst predictions drawn from a model of the measurement process (Tal [2012, 2016]) and Hugo Beauchemin's discussion of systematic uncertainty assessment as an essential component of measurement needed to determine the sensitivity of measurement results in HEP to the physical phenomena or hypotheses for which they might serve as evidence (Beauchemin [2015]).

Section 2 discusses the concept of systematic uncertainty, surveying the ways in which uncertainty has been distinguished from error, and systematic uncertainty from statistical uncertainty. Section 3 uses an example of a typical HEP measurement to illustrate the complexities and importance of systematic uncertainty and section 4 introduces some of the debates among particle physicists regarding the appropriate statistical framework for the estimation of systematic uncertainty. Section 5 outlines the secure evidence framework employed in the analysis. I present my argument for viewing systematic uncertainty estimation as a kind of robustness analysis aimed at establishing empirical value in section 6 and discuss the tension between security and sensitivity in section 7. A brief summary appears in section 8.

As an appendix, I present a discussion of an illuminating example from recent particle physics: the ATLAS collaboration's measurement of the $t\bar{t}$ production cross section from single lepton decays. The case illustrates the proposed analysis of systematic uncertainty assessment, highlights some subtleties in its application, and exemplifies the pervasive character of modelling and simulation in systematic uncertainty estimation in HEP.

2 Systematic Uncertainty: The Very Idea

To facilitate better conceptual understanding, we can begin by clarifying our terminology, with some help from discussions in the field of metrology, the science of measurement and its

applications. Above I referred to both error and uncertainty as being distinguished into systematic and statistical categories. The two terms have distinct histories of usage in science. The scientific analysis of error dates to the seventeenth century, while 'the concept of uncertainty as a quantifiable attribute is relatively new in the history of measurement' (Boumans and Hon [2014], p. 7).

In practice, particle physicists have not always been careful to distinguish between error and uncertainty. Recent papers from the CMS and ATLAS collaborations focus their discussions on uncertainty rather than error, although usage is not perfectly uniform in this regard.

Metrologists, by contrast, have articulated clear distinctions between error and uncertainty, as befits the science whose concern is the very act of measurement. Yet the usefulness and definition of these terms remain matters of debate among metrologists, whose Joint Committee for Guides in Metrology (JCGM) publishes the 'Guide to the Expression of Uncertainty in Measurement' (GUM) (JCGM Working Group I [2008]) and the 'International Vocabulary of Metrology' (VIM) (JCGM Working Group II [2012]). Those debates have turned significantly on the question of the definition of error as articulated in these canonical texts, particularly insofar as that definition appeals to an unobservable, even 'metaphysical' concept of the 'true value' of the measurand (JCGM Working Group I [2008], p. 36).

Some metrologists have defended the importance of retaining a concept of error defined in terms of a true value or 'target value' (Mari and Giordani [2014]; Willink [2013]; Rabinovich [2007]). It is not the purpose of this paper to debate these issues. For the sake of clarity, I will adopt the terminology of Willink ([2013]) and understand a measurement to be a process whereby one obtains a numerical estimate x (the measurement result or measurement estimate) of the target value θ of the measurand. This usage allows for the straightforward definition of measurement error as the difference between the measurement result x and the target value θ.

We may then regard statistical and systematic error as components of the overall measurement error, and turn our attention to how they are to be distinguished.

Following the GUM, we could define the former (also called random error) as the difference between the measurement result x and 'the mean that would result from an infinite number of measurements of the same measurand carried out under repeatability conditions' (JCGM

Working Group I [2008], p. 37). Note that if the measurement procedure is itself biased, then the latter quantity, that which would emerge as the mean measurement result in the long-run limit, will not be equal to the target value. The systematic error, then, is the difference between these two values.

One may approach the concept of systematic error by imagining a measurement process in which it is absent. For such a process, the 'mean that would result from an infinite number of measurements of the same measurand carried out under repeatability conditions' simply would be the target value. Systematic error, then, is a component of measurement error 'that in replicate measurements remains constant or varies in a predictable way' (JCGM Working Group II [2012], p. 22; see also Willink [2013]) and therefore does not disappear in the long run.

Eran Tal, in proposing a model-based account of measurement, has noted an important limitation of this conceptualization of measurement error, which is that it obscures the central role played by the model of the measurement process. The JCGM's appeal to an infinite number of measurements carried out under repeatability conditions relies implicitly on an unspecified standard as to what constitutes a repetition of a given measurement of a measurand. A model of the measurement process not only supplies that standard, it serves to articulate the quantity measured by a given process and thus helps to specify what kinds of measurement outcomes constitute errors. To make these roles of the model explicit, Tal proposes a 'methodological' definition of systematic error as 'a discrepancy whose expected value is nonzero between the anticipated or standard value of a quantity and an estimate of that value based on a model of a measurement process' (Tal [2012], p. 57). The notion of a true value or a target value of the measurand has been supplanted here by an 'anticipated or standard value' that must be ascertained through a calibration process, which in turn is understood as a process of modelling the measuring system. Tal's emphasis on the process by which the error is estimated renders the concept a methodological one, but not yet a purely epistemic concept, for which Tal reserves the term 'uncertainty' (Tal [2012], p. 30).

For purposes of its quantitative treatment, the GUM offers the following definition of uncertainty: 'parameter, associated with the result of a measurement, that characterizes the

dispersion of the values that could reasonably be attributed to the measurand' (JCGM Working Group II [2012], p. 22). This definition clearly marks uncertainty as something potentially quantifiable, but also as something epistemic, requiring consideration of some kind of standard of reasonable attribution. What the GUM's definition does not do, however, is to provide guidance for interpreting this notion of reasonability. Neither does it provide clear guidance in understanding how to characterize the distinction between evaluations of statistical uncertainty and systematic uncertainty. Rather, the GUM eschews this distinction in favor of a purely methodological distinction between Type A uncertainty and Type B uncertainty, based on the method by which uncertainty is evaluated. An evaluation of uncertainty 'by the statistical analysis of series of observations' is Type A. Any other means of evaluation is classified as Type B. This raises a problem for the argument of this paper: Metrologists themselves seem to have concluded that there is no epistemological distinction between statistical and systematic uncertainty, but only a methodological difference. Such an approach renders otiose this paper's attempt to explain the epistemic significance of the evaluation of systematic uncertainty in contrast to statistical uncertainty. There is no epistemic significance to be explained.

As a preliminary observation regarding this problem, it should be noted that experimental HEP has not adopted the GUM's proposal to eliminate the distinction between statistical and systematic uncertainty. This ongoing scientific practice deserves at least to be noted and understood. More fundamentally, the GUM's strictly methodological approach leaves us with the question of why uncertainty itself (which is clearly epistemic by the GUM's own definition given in the previous paragraph) should have components that require evaluation by such different means. One could therefore understand the argument of this paper itself as a defense of an epistemic understanding of the distinction between these two components of uncertainty: The difference in the methods used to evaluate them rests on an epistemological distinction between the quantities being evaluated. To anticipate the argument to come, that difference is that statistical uncertainty is a characteristic of the distribution of outputs of a single model of a measurement process, while systematic uncertainty characterizes the dispersion of measurement results that are obtained when the investigators implement

epistemically possible variations to that model.

3 Systematic Uncertainty in High Energy Physics

To better appreciate the distinction between statistical and systematic uncertainty, and to think more concretely about the epistemic work accomplished by the evaluation of systematic uncertainty, let us consider the disciplinary practices for dealing with systematic uncertainty as it arises in measurements in HEP. We begin with an example.

Measurements of cross sections are a standard part of the experimental program of HEP research groups. The cross section σ quantifies the probability of an interaction process yielding a certain outcome, such as the interaction of two protons in an LHC collision event yielding a top quark – anti-top quark pair (the $t\bar{t}$ production cross section). At its crudest level, such a measurement is simply a matter of counting how many times N, in a given data set, a $t\bar{t}$ pair was produced, and then dividing N by the number L of collision events that occurred during data collection (the latter number particle physicists call the luminosity). We might call this the 'fantasyland' approach to measuring the $t\bar{t}$ production cross section: $\hat{\sigma}_{t\bar{t}} = \frac{N}{L}$. (The 'hat' notation $\hat{\sigma}$ indicates our concern with an equation for determining the value of an estimate of the physical quantity σ.)

Reality intervenes in several ways to drive the physicist out of fantasyland:

(1) $t\bar{t}$ pairs are not directly observable in particle physics data, but must be identified via the identification of their decay products. These products are also not directly observable but must be inferred from the satisfaction of data selection criteria ('cuts'). Events that satisfy these criteria are candidates for being events containing $t\bar{t}$ pairs.

(2) $t\bar{t}$ candidate events may not contain actual $t\bar{t}$ pairs; in other words, they may not be signal events. Other particle processes can produce data that are indistinguishable from $t\bar{t}$ decay events. These events are background. It is the nature of background candidate events that they cannot, given the cuts in terms of which candidates are defined, be distinguished from the signal candidate events that one is aiming to capture; one can only estimate the number to be expected N_b and subtract it from the total number of candidate events observed N_c.

(3) Just as some events that do not contain $t\bar{t}$ pairs will get counted as candidate events,

some events that do contain $t\bar{t}$ pairs will not get thus counted. This problem has two facets.

(3a) The $t\bar{t}$ production cross section as a theoretical quantity might be thought of in terms of an idealized experiment in which every $t\bar{t}$ pair created would be subject to detection in an ideal detector with no gaps in its coverage. Since actual detectors do not have the ability to detect every $t\bar{t}$ event, this limitation of the detector must be taken into account by estimating the acceptance A.

(3b) The production and decay of $t\bar{t}$ pairs are stochastic processes and the resulting decay products will exhibit a distribution of properties. The cuts that are applied to reduce background events will have some probability of eliminating signal events. The solution to this is to estimate the efficiency ϵ of the cuts: the fraction of signal events that will be selected by the cuts.

(4) The physical properties of the elements in a collision event are not perfectly recorded by the detector. Candidate events are defined in terms of quantitative features of the physical processes of particle production and decay. For example, top quarks decay nearly always to a W boson and a b quark. A $t\bar{t}$ pair will therefore result in two W bosons, each of which in turn can decay either into a quark-antiquark pair or into a lepton-neutrino pair. To identify $t\bar{t}$ candidate events via the decay mode in which one of the W bosons decays to a muon (μ) and a muon neutrino (ν_μ), physicists might impose a cut that requires the event to include a muon with a transverse momentum p_T^μ of at least 10 GeV, in order to discriminate against background processes that produce muons with smaller transverse momenta. Whether a given event satisfies this criterion or not depends on a measurement output of the relevant part of the detector, and this measuring device has a finite resolution, meaning that an event that satisfies the requirement $p_T^\mu > 10$ GeV might not in fact include a muon with such a large transverse momentum. Conversely, an event might fail the p_T^μ cut even though it does in fact include a such a muon.

These detector resolution effects require physicists who wish to calculate the $t\bar{t}$ production cross section to base that calculation not simply on the number of candidate events as determined from the comparison of detector outputs to data selection criteria, but on the inferred physical characteristics of the events taking into account detector resolution effects.

This process, called unfolding, requires applying a transformation matrix (estimated by means of simulation) to the detector outputs. Unfolding is a matter of inferring from the detector outputs (via the transformation matrix) the underlying distribution in the sample events, to which the cuts are then applied (Beauchemin [2015], p. 23).

(5) Finally, the luminosity is also not a quantity that is susceptible to direct determination, since distinct events might not get discriminated by the detector, a single event might mistakenly get counted as two distinct events, and some events might be missed altogether. The luminosity must therefore be estimated.

We have thus gone from the fantasyland calculation $\hat{\sigma}_{t\bar{t}} = \frac{N}{L}$ to the physicists' calculation

$$\hat{\sigma}_{t\bar{t}} = \frac{N_c - N_b}{\epsilon A L}. \tag{3.1}$$

This calculation is not merely more complex than the fantasyland calculation. Every quantity involved in it is the outcome of an inference from a mixture of theory, simulation, and data (from the current experiment or from other experiments).[2] Each has its own sources of uncertainty that the careful physicist is obliged to take into account.

4 Methdological Debates in High Energy Physics

But how ought one to take these uncertainties into account? HEP lacks a clear consensus.

Discussions about the conceptualization of error and uncertainty, and about their classification into categories such as statistical and systematic, are inseparably bound up with debates regarding the statistical framework in which these quantities should be estimated and expressed. When discussion focuses on statistical error alone, the applicability of a strictly frequentist conception of probability stirs up no significant controversy. One can clearly incorporate into one's model of an experiment or measuring device a distribution function representing the relative frequency with which the experiment or device would indicate a range of output values (results) for a given value of the measurand. Such a model, which will

[2]This discussion affirms Wendy Parker's recent argument that computer simulations can be 'embedded' in measurement practices (Parker [2015]).

inevitably involve some idealization, can be warranted by a chain of calibrations. Indeed, as argued by Tal ([2012]), the warranting of inferences from measurement results in general requires such idealization. One can then incorporate this distribution of measurement errors into one's account of the measurement's impact on the uncertainty regarding the value of the measurand.

Systematic errors cannot be treated in this same straightforward manner. Consider the paradigmatic example of systematic error: a biased measuring device. Suppose that a badly-constructed ruler for measuring length systematically adds 0.5 cm whenever one measures a 10.0 cm length. Repeated measurements of a given 10.0 cm standard length would produce results that cluster according to some distribution around the expectation value 10.5 cm. The difference between the 10.0 cm standard length and the expectation value of 10.5 cm just is the systematic error on such measurements. If we know that this bias is present, we can eliminate the error by correction.

The problem of the estimation of systematic uncertainty arises precisely when one cannot apply the correction strategy because the magnitude of the error is unknown. The investigator knows that a systematic error might be present, and the problem is to give reasonable bounds on its possible magnitude. To this problem the notion of a frequentist probability distribution no longer has an obvious direct applicability; the error is either systematically present (and with some particular, but unknown, magnitude), or it is not.[3]

As a consequence, investigators employing frequentist statistics to evaluate statistical uncertainty report systematic uncertainty separately, as particle physicists typically do. The quantities denoted 'statistical' and 'systematic' in a statement such as '$\sigma_{t\bar{t}} = 187 \pm 11(\text{stat.})^{+18}_{-17}(\text{syst.})$ pb' are conceptually heterogeneous. Combining them into a

[3]See, however, Cranmer [2003] for a step towards a strictly frequentist approach. Willink ([2013]) also argues that a frequentist construal of systematic uncertainty bounds is applicable for the consumer, rather than the producer, of measurement results, if one adopts an enlarged view of the measurement process to include the 'background steps' that introduce systematic errors, so that any one measurement result can be regarded as having been drawn from a population that includes a variety of different background steps.

single quantity and calling it the 'total uncertainty' is problematic.

One response to this problem is to adopt the Bayesian conception of probability as a measure of degree of belief. Such a shift from frequentist to Bayesian probabilities is advocated in the VIM as part of the shift from an Error Approach to an Uncertainty Approach (JCGM Working Group II [2012]). The expectation value of a Bayesian probability distribution is no longer understood as the mean in the long-run limit, but the average of all possible measurement results weighted by how strongly the investigator believes that a given result will obtain, when applied to a given measurand. A putative advantage of this Bayesian approach is that it allows for the straightforward synthesis of statistical and systematic uncertainties into a single quantity. Both Sinervo ([2003]) and the JCGM ([2008], p. 57) cite this as a point in favor of a Bayesian understanding of the probabilities in a quantitative treatment of systematic uncertainty.

Adopting a Bayesian approach comes with well-known difficulties, however, also acknowledged by Sinervo ([2003]; see also Barlow [2002]). Investigators must provide a prior distribution for each parameter that contributes to the systematic uncertainty in a given measurement. Just what the constraints on such prior distributions ought to be (aside from coherence) is very unclear.

A third approach to the problem employs a hybrid of Bayesian and frequentist techniques. The Cousins–Highland method relies on a calculation that takes a frequentist probability distribution (giving rise to the statistical error) and 'smears' it out by applying a Bayesian probability distribution to whatever parameters that distribution might depend on that are sources of systematic uncertainty (Cousins and Highland [1992]).

The basic idea is this: suppose that one has a set of observations $x_i, i = 1, \ldots, n$, distributed according to $p(x|\theta)$, and that the data $\{x_i\}$ are to be used to make inferences about the parameter θ. Now, suppose that such inferences require assumptions about the value of λ, an additional parameter, the value of which is subject to some uncertainty. The hybrid method involves introducing a prior distribution, $\pi(\lambda)$, to enable the calculation of a modified probability distribution $p_{CH}(x|\theta) = \int p(x|\theta, \lambda)\pi(\lambda)d\lambda$, which then becomes the basis for statistical inferences (Sinervo [2003], p. 128).

The Cousins-Highland hybrid approach yields, as critics have noted (Cranmer [2003]; Sinervo [2003]), neither a coherent Bayesian nor a coherent frequentist conception. The statistical distribution $p(x|\theta)$ is intended to be frequentist, but the prior distribution on λ has no truly frequentist significance, leaving the modified distribution $p_{CH}(x|\theta)$ without any coherent probability interpretation. Cousins and Highland defend the approach on the grounds that it adheres as closely as possible to a frequentist approach while avoiding 'physically unacceptable' consequences of a 'consistently classical' approach, specifically, that a stricter upper limit may sometimes be derivable from an experiment that has a larger systematic uncertainty (Cousins [1992, 1995]).[4]

The discussion thus far serves to illustrate some of the ways in which the conceptual underpinnings of systematic uncertainty estimation remain unsettled. The conceptual disorder not only poses an intellectual problem, but contributes to ongoing confusion and controversy over the appropriate methodology for estimating uncertainty. Moreover, the problems spill over into the use of uncertainty bounds when determining the compatibility of one measurement result with another, or with a theoretical prediction. The problem then arises as to how statistical and systematic uncertainties will be combined. It is common to add them in quadrature, for example when using a χ^2 fit test, but doing so introduces distributional assumptions that may be unwarranted, and that may not reflect the manner in which the systematic uncertainty was in fact determined in the first place. One could simply add the two uncertainty components in a linear manner, but this would in many cases significantly and unnecessarily reduce the sensitivity of the measurement results. HEP currently lacks a consensus regarding a satisfactory methodology for treating systematic uncertainty.

Without attempting to resolve such thorny methodological disputes at once, we can progress towards a more satisfactory treatment of uncertainty estimation by first grasping

[4]This result may occur when the uncertainty bounds of an imprecise estimate of a parameter happen to extend into an 'unphysical' region. For example, in estimating the mass of a neutrino one might produce an estimate that extends significantly into negative numbers. The uncertainty interval as a whole might be large, but the upper limit might be very close to zero (James and Roos [1991]).

more clearly the epistemic work such estimates seek to accomplish.

I will argue that by working within the secure evidence framework, we can at least partially assimilate the epistemic work accomplished by systematic uncertainty estimation to that accomplished by robustness analysis. That systematic uncertainty estimation provides a means for investigating the robustness of a measurement result has already been argued in an illuminating essay by Beauchemin ([2015]). By quantifying both the uncertainty of a measurement result and its sensitivity to the phenomena under investigation, he argues, scientists can quantify the scientific value of a measurement result and provide criteria for minimizing the circularity that arises from the theory-laden aspect of measurement.

Here I aim to build upon the insights of Beauchemin by providing a broader epistemological framework for understanding what is achieved through robustness analysis in the context of estimating systematic uncertainty. The secure evidence framework I invoke involves no explicit commitment to either frequentist or Bayesian statistical frameworks. The relevant modality in the framework is possibility, considered to be conceptually prior to probability, insofar as no probability function can be specified without specifying a space of objects (whether events, propositions, or sentences) over which the function is to be defined.

5 The Secure Evidence Framework

Here I explain the secure evidence framework that will be adopted to analyze these issues (Staley [2004, 2011, 2012, 2014]). On the one hand, we might wish to think of the evidence or support for a hypothesis provided by the outcome of a test of that hypothesis in objective terms, such that facts about the epistemic situation of the investigator are irrelevant. On the other hand, it seems quite plausible that evaluating the claims that an investigator makes about such evidential relations may require determining what kinds of errors investigators are in a position to justifiably regard as having been eliminated, which does depend on their epistemic situation. The secure evidence framework provides a set of concepts for understanding the relationship between the situation of an epistemic agent (either individual or corporate) and the objective evidential relationships that obtain between the outcomes of tests and the hypotheses that are tested. This account relies centrally on a concern with possible errors, and

explicitly understands the relevant modality for possible errors to be epistemic.

An epistemic agent who evaluates the evidential bearing of some body of data $\mathbf{x}_0$ with regard to a hypothesis H must also consider the possibilities for error in the evaluation thus generated. This is the 'critical mode' of evidential evaluation. Evidential judgments rely on premises, and errors in the premises of such a judgment may result in errors in the conclusion. A responsible evaluation of evidence therefore requires consideration of the ways the world might be, such that a putative evidential judgment would be incorrect. The evaluator must reflect on what, among the propositions relevant to the judgment in question, may safely be regarded as known, and what propositions must be regarded as assumed, but possibly incorrect.

Such possibilities of error are here regarded as epistemic possibilities. Often, when one makes a statement in the indicative that something might be the case, one is expressing an epistemic possibility, with what must be the case functioning as its dual expressing epistemic necessity (Kratzer [1977]). An expression roughly picking out the same modality (at least for the singular first-person case) is 'for all I know', as in 'for all I know there is still some ice cream in the freezer'. Theorists have offered a range of views regarding the semantics of epistemic modality (see Egan and Weatherson [2011]), with various versions of contextualism and relativism among the contending positions. For our purposes we need only note that what is epistemically possible for an epistemic agent does depend on that agent's knowledge state and that when an agent acquires new knowledge it follows that some state of affairs that was previously epistemically possible for that agent ceases to be so.

The epistemic possibilities that are relevant to the critical mode of evidential judgment are error scenarios, which are to be understood as follows: Suppose that P is some proposition, S is an epistemic agent considering the assertion of P, whose epistemic situation (her situation regarding what she knows, is able to infer from what she knows, and her access to information) is K. Then, a way the world (or some relevant part of the world) might be, relative to K, such that P is false, we will call an error scenario for P relative to K.

Of special importance for this discussion are error scenarios for evidence claims, where EC is an evidence claim if it is a statement expressing a proposition of the form 'data $\mathbf{x}$ from test

T are evidence for hypothesis H' (such hypotheses may include statements about the value of some measurand in a measurement process, and it is here assumed that measurement procedures and hypothesis testing are epistemologically susceptible to the same analysis).

Suppose that, relative to a certain epistemic situation K, there is a set of scenarios that are epistemically possible, and call that set Ω_0. If proposition P is true in every scenario in the range Ω_0, then P is fully secure relative to K. If P is true across some more limited portion Ω_1 of Ω_0 ($\Omega_1 \subseteq \Omega_0$), then P is secure throughout Ω_1.

To put this notion more intuitively, then, a proposition is secure for an epistemic agent just insofar as that proposition remains true, whatever might be the case for that agent. Thus defined, security is applicable to any proposition, but the application of interest here is to evidence claims.

Note that inquirers might never be called upon to quantify the degree of security of any of their inferences. The methodologically significant concept is not security as such, but the securing of evidence, which is to say, the pursuit of strategies that increase or assess the relative security of an evidence claim.

We may distinguish two strategies for making inferences more secure: the weakening of evidential conclusions to render them immune to otherwise threatening error scenarios and the strengthening of premises, in which additional information is gathered that rules out previously threatening error scenarios. Robustness analysis constitutes a strategy for assessing the security of an evidence claim by investigating classes of potential error scenarios to determine which scenarios are and which are not compatible with a given evidential conclusion (Staley [2014]).

The present analysis aims to show how the consideration and evaluation of systematic uncertainty constitutes an application of the weakening strategy under the guidance of robustness analysis. The focus in the secure evidence framework is on evidence claims; here the relevant evidence claims assert that data collected in the measurement process are evidence in support of the measurement result. I will sometimes speak of securing the measurement result, but this should be understood as shorthand for securing the evidence in support of the measurement result. (Experimentalists in HEP do not typically report their

measurement results in the form of explicit statements of the form 'data **x** support the measurement result $\mu = \mu_0 \pm a \pm b$'. They simply state their result. Clearly, however, they would not be in a position to state the measurement result in the absence of such evidential support, so I treat the corresponding evidence claim as implied.)

6 Systematic Uncertainty Assessment as Robustness Analysis

To see how the evaluation of systematic uncertainty can be viewed as a variety of robustness analysis, consider again the example of measuring the $t\bar{t}$ production cross section in proton-proton collisions. Recall how equation 3.1 relates an estimate of that quantity to other empirically accessible quantities:

$$\hat{\sigma}_{t\bar{t}} = \frac{N_c - N_b}{\epsilon A L}.$$

This equation is a premise of an argument for a conclusion about the value of $\sigma_{t\bar{t}}$, as are statements attributing values to each of the variables in the equation.

If we take the conclusion of such an argument to be the attribution of some definite value $\sigma_{t\bar{t}} = \sigma_{t\bar{t}_0}$ (for example, $\sigma_{t\bar{t}} = 187$pb), then the premises would fail to provide cogent support for the conclusion. The data are finite. Were we to repeat the exact same experiment, it is very probable that a different value assignment would result. It is not clear that such a statement should be considered an experimental conclusion at all. It appears to be a statement about the value of the quantity $\sigma_{t\bar{t}}$, but its usefulness for empirical inquiry is severely limited by the fact that the comparison of any two such point-value determinations is almost certain to yield the result that they disagree, regardless of the actual (target) value of $\sigma_{t\bar{t}}$. To restore cogency and empirical value, it is necessary to replace the conclusion $\sigma_{t\bar{t}} = \sigma_{t\bar{t}_0}$ with one stating $\sigma_{t\bar{t}} = \sigma_{t\bar{t}_0} \pm \delta$. For reasons that will become apparent, let us call this the unsecured conclusion.

The addendum '$\pm\delta$' expresses the statistical uncertainty that is a function of the number of candidate events, N_c. This uncertainty reflects the fact that, even were all of the input quantities in equation 1 perfectly known, repeating the measurement will not generally yield the same value for the esimator $\hat{\sigma}_{t\bar{t}}$. But it will also be necessary to report the systematic uncertainty resulting from the fact that knowledge of the acceptance, efficiency, background,

luminosity, and unfolding matrix used to determine the number of candidate events is imperfect. That conclusion, which includes a statement of statistical uncertainty, but adds the results of an assessment of systematic uncertainty, we can call the secured conclusion.

Importantly, lack of knowledge enters into statistical and systematic uncertainties in quite different ways. In the case of statistical uncertainty, the relevant lack of knowledge concerns the exact value of the quantity $\sigma_{t\bar{t}}$, the measured value of which is reported in the conclusion. The aim of the inquiry is to reduce this uncertainty, and knowledge of the quantities on the right hand side of equation 3.1 is a means toward the achievement of this aim. The statistical uncertainty that remains once those means have been deployed is a consequence of the fact that any finite number of observations yields only partial information about the value of $\sigma_{t\bar{t}}$. For the purposes of assessing statistical uncertainty, however, the premises in the argument for this (unsecured) conclusion are assumed to be determinately and completely known. We take ourselves to know how many candidate events were counted, for example, and as long as that number is finite, there will be some corresponding statistical uncertainty on the estimate of $\sigma_{t\bar{t}}$.

In other words, for the purposes of assessing statistical uncertainty, we concern ourselves only with the possibility of errors in the conclusion of the argument, errors that take the form of assigning incorrect values to the measurand, not because of an error in the premises, but as a result of the variance of the estimator. For this purpose, the model of the measuring process is taken to be adequate and the premises of the argument are assumed to be free of error.

Systematic uncertainty may then be regarded as arising from the consideration that in fact the premises are not determinately and completely known. Physicists are not in a position to know that a premise attributing a definite value to, say, ϵ is true. To derive an estimate of $\sigma_{t\bar{t}}$, some value must be assigned, but, having made such an assignment, investigators must confront the fact that other value assignments are compatible with what they know about the detector and the physical processes involved in the experiment.

The assessment of systematic uncertainty tackles this problem of incomplete knowledge by, in some way, exploring the extent to which varying the assumed values asserted in the premises, within the bounds of what is possible given the investigators' knowledge, makes a difference to the conclusion drawn regarding the value of the measured quantity. This

effectively generates an ensemble of model-based arguments corresponding to the considered range of possible value-assignment premises. By considering such a range of arguments with possibly correct premises, the investigators can then report a weakened (secured) conclusion such as

$$\sigma_{t\bar{t}} = 187 \pm 11(\text{stat.})^{+18}_{-17}(\text{syst.})\text{pb},$$

the correctness of which would be supported by the soundness of any one of the arguments in the ensemble.

From the perspective of the secure evidence framework, such assessments can be regarded as a combination of robustness and weakening strategies aimed at the evidential support that the data provide for the measurement result. Investigators begin with a set of data $x_i, i = 1, \ldots, n$ reporting observations relevant to the measurement of some quantity θ. This derivation depends on assigning values $\lambda_j = \lambda_{j0}, j = 1, \ldots, m$ to each of a set of m imperfectly known parameters necessary for the derivation of an estimate $\hat{\theta}_0$ (with a statistical uncertainty depending on the value of n). This yields the unsecured conclusion. The strategies for securing evidence claims mentioned previously have not yet been applied to it.

Consideration then turns to limitations on the investigators' knowledge of the initial set of model assumptions. Using the robustness analysis strategy, alternate sets of assumptions that are compatible with the investigator's epistemic state are considered, taking the form $\lambda_j = \lambda_{j0} + \varepsilon_j$ for each λ_j and for a range of values of ε_j, depending on the extent to which existing knowledge constrains the possible values of λ_j. This yields a range of derived estimates lying within an interval $\hat{\theta}_0{}^{+\delta_1}_{-\delta_2}$, which then can provide the basis for a logically weakened conclusion incorporating both statistical and systematic uncertainties. The convergence of estimates generated by the ensemble within the interval $\hat{\theta}_0{}^{+\delta_1}_{-\delta_2}$ provides the basis for the robustness of this secured conclusion.

The account just given is misleading in one respect. The evaluation of systematic uncertainty in HEP is typically not a matter of directly varying the input quantities in an equation such as equation 3.1. Instead, physicists look upstream, to the methods and models employed in the determination of those input quantities, and introduce variation there. In some cases, this is a matter of varying the value of some parameter in a model, in other cases it is a

matter of swapping one model for another. What is at issue in such variation is not so much the question of what might be the true value of the parameter or the true model (notions that might be inapplicable in a given case) as which models or parameter values within a model might be adequate for the purposes of the inference that is being undertaken (Parker [2009, 2012]).

This point will be documented in greater detail in the appendix, but a brief illustration from the case discussed there might help to clarify how both variation of the model and variation of parameter values in a model play a role in evaluating systematic uncertainty.

In a measurement of the top quark–anti-top quark pair production cross section, the ATLAS collaboration estimated the acceptance for $t\bar{t}$ events. To estimate that quantity, they simulated the signal events using a Monte Carlo simulation, one of several that physicists have developed over the years. That Monte Carlo simulation might be thought of as a component of the model of ATLAS's measurement process, as it contributes to the generation of a crucial input quantity to equation (1). To see how much systematic uncertainty is due to their choice of simulation, ATLAS swaps out one simulation for another and checks to see how much difference that makes to the calculation of the cross section. This is an example of varying the model by changing one of its components.

Another contribution to the systematic uncertainty on ATLAS's measurement concerned the identification of candidate events as possessing certain physical attributes, such as the presence of at least three 'jets' of hadronic particles with momenta meeting certain threshold requirements. Such determinations are themselves subject to error, however, dependent upon, for example, a 'Jet Energy Scale' (JES). Determining the uncertainty due to the choice of JES involved performing 'pseudo-experiments' on simulated data 'with jet energies scaled up and down according to their uncertainties and the impact on the cross-section was evaluated' (ATLAS [2012d], p. 250). Here the variation is applied to a parameter in a model of the measurement process.

Attention here has focused on the evaluation of systematic uncertainty as involving the use of a weakening strategy. A broader consideration of systematic uncertainty reveals the importance of strengthening strategies as well, since much of the effort of experimentalists in the generation and analysis of data in a measurement experiment is aimed at reducing

systematic uncertainty. Much of the repertoire of good experimental practice in HEP can be understood as aimed at the reduction of both statistical and systematic uncertainty. Collecting and selecting high-quality 'control' data for estimating backgrounds, choosing judiciously the Monte Carlo simulations for modelling signal, calibration and shielding of detector components, and the determination of data selection criteria for defining the class of candidate events: all of these activities and more are carried out with an overriding concern for enabling the measurement to achieve a sufficiently small uncertainty on the ultimate result for the learning aims of the experiment.

The secure evidence framework gives us a new perspective on what is distinctive about the assessment of systematic uncertainty: it involves consideration of what might be the case regarding parameters in the model of the experiment to which values must be assigned in order to derive an estimate of the measured quantity. Systematic uncertainty is thus concerned with possible errors and inadequacies in the premises of an argument based on a model of measurement processes. The determination of systematic uncertainty involves the determination of reasonable bounds on the possibility of errors in those premises and inadequacies in the model of the process. Such analysis rests on an ensemble of completed models, each of which corresponds to a potential error scenario regarding claims about the value of the measured quantity. Only through the consideration of the outputs of such an ensemble can a systematic uncertainty be assessed, whereas claims about statistical uncertainty require only a single completed model.[5]

One might object that the interpretation of systematic uncertainty estimation as a weakening strategy applied to the unsecured conclusion is an artifact of an unwarranted assumption that statistical uncertainty is somehow logically or epistemically prior to systematic uncertainty.[6] According to this objection, one could just as readily treat statistical

[5]As an anonymous referee pointed out, this account calls to mind the use of model ensembles in climate projections, and it may be that the current perspective can be extended to this and other cases (see, for example, Parker [2011]; Vezér [2016]). Significant differences are likely to distinguish the climate science and HEP cases, however; I will not here speculate further on the prospects for extending this analysis to other cases.

[6]I am grateful to an anonymous referee for this very helpful objection.

uncertainty as a weakening of a result incorporating systematic uncertainty.

The idea behind such an approach is that one could use of each of an ensemble of completed models to generate a point estimate of the measurand, resulting in a range of values reflecting systematic uncertainty, and then add the statistical distribution appropriate to each completed model to arrive at the statistical uncertainty. While such an approach might be methodologically feasible, it makes no epistemological sense. The completed models used in the first stage are not candidates for being adequate models of the measurement process. That process does exhibit variance in its outputs, a feature that is missing from models that generate only point estimates. An estimate of systematic uncertainty produced in this way would not express uncertainty as to which of an ensemble of models might be an adequate representation of the measurement process, since all of the models in the ensemble would be known to be inadequate at the outset.

The distinction between statistical and systematic uncertainty estimation corresponds to that between the use mode and the critical mode in model-based evidential reasoning (Staley [2014], pp. 41–42). In use mode, a model of the measurement process is used to generate an estimate of the measurand from the data. The critical mode focuses on uncertainties regarding the adequacy of that model, and the strategy of robustness analysis addresses those uncertainties by varying the model within the bounds of what is epistemically possible. Statistical uncertainties are integral to the use of any single model.

This perspective can shed light on debates over statistical methodology in the evaluation of systematic uncertainty. For example, consider the the Bayesian-frequentist hybrid approach developed by Cousins and Highland, discussed above. As mentioned, one objection to this method is that modifying the sampling distribution with a prior probability on the nuisance parameter fails to yield a conceptually coherent probability distribution. From the perspective on systematic uncertainty advanced here, this probability distribution can be useful for the investigator's aims without being susceptible to a coherent probabilistic interpretation. It serves to incorporate consideration of a range of epistemically possible values of the nuisance variable, which allows the resulting distribution to be used to calculate the range of values that the measurand might take with the probability of (say) $p = 0.68$. The probability invoked in

this characterization could be given a frequentist interpretation in the case where there are no nuisance parameters (the model of the measurement process is known to be adequate). The term 'might' indicates that this is not our situation. A range of model possibilities are compatible with out background knowledge. Although there is some definite frequentist distribution of measurement outcomes for each one of those possibilities, the hybrid treatment permits the investigator to use background knowledge so as to give more weight to some of those possibilities than to others, such that the resulting uncertainty interval reflects both the statistical behaviour of the measurement process and the physicists' judgment of the knowledge state employed in assessing the possibilities of error in the characterization of that measurement process. The loss of interpretive clarity is not to be taken lightly, but that cost must be weighed in the context of the aims of evaluating systematic uncertainty.

Such a perspective also helps us to understand why systematic uncertainty is important in science and should be important in the philosophy of science. As noted in the introduction and documented in the appendix, while statistical uncertainties are simply reported, a good assessment of systematic uncertainties involves careful consideration of a wide range of factors that are relevant to the conclusion being drawn from the data, as well as careful probing of the limitations on the investigators' prior knowledge of those factors. Whereas a model of the experiment is used to arrive at an unmodulated conclusion, the modulated conclusion depends on a stage of critical assessment of that model, achieved through a process of robustness analysis.

7 Security and Sensitivity

The empirical value of a measurement result depends not only on its security, however, but also on its sensitivity to physical phenomena of interest, as explained by Beauchemin ([2015]). Sensitivity is concerned with the ability of a measurement result to inform our answers to substantive physics questions.

In Beauchemin's discussion, the sensitivity of a measurement result depends on the comparison of '(1) the difference in the values of a given observable when calculated with different theoretical assumptions to be tested with the measurement; and (2) the uncertainty on

the measurement result' (Beauchemin [2015], p. 29). Only if the systematic uncertainty is sufficiently small in comparison to the differences between different theoretically calculated quantity values can the measurement result be used to discriminate empirically between the competing theoretical assumptions.

Sensitivity considerations thus reveal the cost of applying the weakening strategy to enhance the security of a measurement result: One can weaken the conclusion of an argument for such a result by enlarging the systematic uncertainty bounds applied to it, thus achieving a conclusion that is secure across a broader range of epistemically possible scenarios. However, if the systematic uncertainties thus reported exceed the differences between values of the measurand yielded by the theoretical assumptions to be tested amongst, then the level of security achieved renders the measurement result empirically worthless relative to that testing aim. It becomes 'compatible within uncertainty' with all the theories among which one aims to choose.

One significant problem should be noted before closing, though it deserves much greater consideration in future work. On the secure evidence approach, evaluating systematic uncertainty involves the determination of reasonable bounds on the possibility of errors in the premises used, and inadequacies in the model of the process. Which possibilities deserve consideration? What criteria ought to be considered when determining standards of reasonableness on such bounds? The previously mentioned debates over methodology can be thought of as debates over the best way to approach such questions. Bayesian methods provide investigators with the freedom to choose a prior distribution that reflects their judgment about the probability of alternative values of parameters within a model, while frequentist approaches prioritize the communication of the operating characteristics (in terms of error probabilities) of the measurement procedure used as a function of those parameters. In practice, techniques such as switching Monte Carlo generators and recording the difference it makes to the estimate of the measurand amount to judgments about which other methods one might reasonably have used in carrying out the experiment.

8 Conclusion

Given the scant attention that the evaluation of systematic uncertainty has received in the philosophical literature, it is appropriate that conclusions drawn at this stage of inquiry should be provisional and exploratory in nature. I have proposed that, while many questions regarding the assessment of systematic uncertainty remain unresolved, a good first step towards a philosophical appreciation of this practice is to regard it as a kind of robustness analysis. The secure evidence framework provides a context for understanding robustness analysis in general that also supports this identification of systematic uncertainty assessment as a kind of robustness analysis. That a coherence can thereby be established between this view of systematic uncertainty assessment and Tal's model-based account of measurement is an additional virtue of the present account.

Further discussion is, of course, needed. Of particular importance is to engage the philosophical aspects of the debate over statistical methodology in systematic uncertainty estimation. Gaining clarity about the aims of systematic uncertainty estimation is crucial for this purpose. I have argued here for the view that the aim of this practice is to enable the construction of model-based arguments for secure conclusions regarding the value of a measurand while enabling the investigator to assess the empirical value of that measurement result. In the previous section I suggested a way in which this aim could help to understand the relevance of the Bayesian-frequentist hybrid approach in spite of conceptual problems in its interpretation. That suggestion remains provisional, and subject to further consideration of fully frequentist and fully Bayesian alternatives.

Appendix A Measuring the $t\bar{t}$ Production Cross Section

Because assessments of systematic uncertainty are considered essential to the publication of any experimental result in particle physics, the choice of an example to illustrate the practice is largely arbitrary. Here I present a recent example from the ATLAS group at the Large Hadron Collider (LHC) at CERN, which also illustrates the pervasive character of modelling and simulation in the analysis of experimental data in contemporary HEP (Morrison [2015]).

The cross section for a given state quantifies the rate at which that state is produced out of some particle process. Cross sections serve as crucial parameters in the Standard Model, and, especially in the case of the top quark, provide important constraints on the viability of numerous Beyond–Standard Model theories as well. Both ATLAS and CMS, its neighbor at the LHC, have published a number of measurements of the production cross section for both top–anti top ($t\bar{t}$) pairs (ATLAS [2012b, 2012d, 2012c]; CMS [2011, 2013]) and for single top quarks (ATLAS [2012a]). The present discussion focuses on a measurement of the top quark pair production cross section based on a search for top quark decays indicated by a single high momentum lepton (electron or muon) and jets produced by strong interaction processes characterized by Quantum Chromodynamics (QCD).

Let us begin with the result that ATLAS reports:[7]

$$\sigma_{t\bar{t}} = 187 \pm 11(\text{stat.})^{+18}_{-17}(\text{syst.})\text{pb} \tag{A.1}$$

We have already seen the essential logic of such a measurement: One seeks to estimate the rate at which $t\bar{t}$ pairs are produced from the number of $t\bar{t}$ candidate events in the data. Making that inference, however, demands estimates of the quantities in equation 3.1, and the accurate estimation of those quantities demands skill and expert judgment.

The complete estimate of uncertainty in this case draws on more considerations than I can address in a brief discussion (see figure 1), but the following should suffice to communicate the nature of the problem.

[7]The paper reports two cross section estimates using different techniques. The second estimate ($\sigma_{t\bar{t}} = 173 \pm 17(\text{stat.})^{+18}_{-16}(\text{syst.})$ pb) does not employ a technique for tagging jets containing b quarks. ATLAS reports the systematic uncertainty due to the estimate of luminosity separately, adding another ±6 pb to the measurements from each method. Both results, ATLAS states, agree with one another and with QCD calculations, but the method using b-tagging 'has a better a priori sensitivity and constitutes the main result of this Letter' (ATLAS [2012d], p. 244).

Method	Untagged		Tagged	
Statistical Error (%)	+10.1	−10.1	+5.8	−5.7
Object selection (%)				
JES and jet energy resolution	+4.1	−5.4	+3.9	−2.9
Lepton reconstruction, identification and trigger	+1.7	−1.6	+2.1	−1.8
Background modelling (%)				
Multijet shape	+3.5	−3.5	+0.8	−0.8
Multijet normalisation	+1.1	−1.2	*a/i*	
Small backgrounds norm.	+0.6	−0.6	*a/i*	
W + jets shape	+3.9	−3.9	+1.0	−1.0
W + jets heavy-flavour content	*n/a*		+2.7	−2.4
b-tagging calibration	*n/a*		+4.1	−3.8
$t\bar{t}$ signal modelling (%)				
ISR/FSR	+6.3	−2.1	+5.2	−5.2
NLO generator	+3.3	−3.3	+4.2	−4.2
Hadronisation	+2.1	−2.1	+0.4	−0.4
PDF	+1.8	−1.8	+1.5	−1.5
Others (%)				
Simulation of pile-up	+1.2	−1.2	< 0.1	
Template statistics	+1.3	−1.3	+1.1	−1.1
Systematic Error (%)	+10.5	−9.4	+9.7	−9.0

Figure 1: Table of statistical and systematic uncertainties for two different analyses, one requiring events to include a jet from a b-quark ('tagged') and one without that requirement ('untagged'). Note that everything below the first line ('statistical error') is a contribution to the systematic uncertainty. The total systematic uncertainty is calculated by adding the individual contributions in quadrature. From ATLAS [2012b], p. 250.

Consider first the estimation of signal acceptance and efficiency. Estimating these quantities requires consideration of characteristics of the detector, drawing on the engineering knowledge of the detector's design and construction as well as experimental knowledge of its performing characteristics. This background knowledge forms the basis for a computer simulation of the detector itself. Estimating signal acceptance and efficiency also involves the knowledge of the characteristics of the signal itself: how are $t\bar{t}$ pairs produced in the proton-proton collisions generated at the LHC, and how do they behave once they have been produced? To model the $t\bar{t}$ signal, ATLAS uses a variety of simulation models and a variety of parameter values within those models. It is this variation of modelling assumptions and their role in the production of systematic uncertainty estimates that I wish to emphasize.

To simulate the production of $t\bar{t}$ pairs in the collider environment, numerous stochastic QCD processes must be estimated, none of which can be calculated exactly from theory. The underlying event is the interaction between colliding high energy protons. Particles involved in the collisions and their subsequent decay products also emit QCD radiation, which is relevant to the calculation of the probability of various outcomes. Both the Initial State Radiation (ISR, prior to the beam collision) and Final State Radiation (FSR, subsequent to the collision) must be modeled as well. Finally, quarks and gluons that are produced in these processes become hadrons (bound states of quarks with other quarks), a process known as hadronization.

To estimate the rate at which $t\bar{t}$ pairs produced in $\sqrt{s} = 7$ TeV proton-proton collisions will qualify as candidate events, ATLAS must simulate these physical processes using a collection of computer simulations that have been developed over the years by physicists. The simulations are based on theoretical principles and constrained by existing data from previous particle physics endeavors. They use the Monte Carlo method of generating approximate solutions to equations that cannot be solved analytically.

ATLAS relies primarily on the Herwig (Hadron Emission Reactions With Interfering Gluons) event generator. To model the further development of a collision event, ATLAS used the event generator MC@NLO (for Monte Carlo at Next-to-Leading-Order). This is a simulation that calculates QCD processes at the level of next-to-leading-order accuracy but also models the parton showers of QCD radiation that result from proton-proton collisions.

To further complicate things, the outcome of a proton-proton collision depends on the way in which the momentum of the proton is distributed among its constituent partons, described probabilistically by the Parton Distribution Function (PDF). So crucial is the judicious choice of PDF in the simulation of particle processes at the LHC that a special LHC working group (PDF4LHC) has devoted its efforts to the formulation of recommendations for the choice of PDF sets for particular LHC analyses (Botje *et al.* [2011]).

'The use of simulated $t\bar{t}$ samples to calculate the signal acceptance gives rise to various sources of systematic uncertainty. These arise from the choice of the event generator and PDF set, and from the modelling of initial and final state radiation' (ATLAS [2012d], p. 245). Evaluation of these uncertainties involves the quantitative assessment of how much difference variations in those assumptions make to the estimate they generate.

In explanation of their approach to this task, ATLAS notes that to evaluate uncertainties due to the 'choice of generator and parton shower model' they compared the results they had obtained using MC@NLO with those obtained using an alternate simulation called POWHEG, using either HERWIG or PYTHIA (an alternate event generator) to model the hadronization process. Yet another generator, called ACERMC, in combination with PYTHIA, is used to assess the uncertainty introduced by ISR/FSR assumptions, 'varying the parameters controlling the ISR/FSR emission by a factor of two up and down' (ATLAS [2012d], p. 245). Finally, to evaluate the 'uncertainty in the PDF set used to generate $t\bar{t}$ samples, ATLAS employed 'a range of current PDF sets' following the procedure recommended by the PDF4LHC working group.

Figure 1 gives the results of these procedures, for two different analysis procedures, one requiring events to include a jet from a *b*-quark ('tagged') and one without that requirement ('untagged'), under the heading '$t\bar{t}$ signal modelling'. That table also tabulates all other sources of systematic uncertainty, yielding totals arrived at by adding the individual contributions in quadrature (that is, total systematic uncertainties are equal to the square root of the sums of the squares of the individual contributions).

One of the categories of systematic uncertainty is 'object selection', under which heading we find the entries 'JES [Jet Energy Scale] and jet energy resolution' and 'Lepton

reconstruction, identification and trigger'. The motivation for these entries concerns the way in which candidate events are defined, which is in terms of the identification of decay products with certain properties. For example, this paper focused on $t\bar{t}$ decays with a single high-momentum lepton (electron or muon) and jets from QCD processes. The implementation of this idea was based on the idea that measurements of energy deposits in the detector could be used to identify a track as resulting from the passage of an electron (or muon), the momentum (transverse to the beam) of which could then be measured, to determine whether they satisfied the threshold requirement of $p_T > 20$ GeV. Only events including such a high-momentum lepton and at least three jets with $p_T > 25$ GeV (and meeting further requirements) could be counted as candidate events.

The identification of an event as including a high-momentum lepton and three high-momentum jets, however, has its own uncertainty, and this is what the 'object selection' uncertainty seeks to quantify. There is always some chance that energy will be deposited in the various detector components in a way that will 'fool' the detector into thinking that a high-p_T electron has passed when it has not. The uncertainty that results from this possibility entails that the number of candidate events itself is to some extent uncertain.

Nonetheless, on the present account, such uncertainty remains systematic in character insofar as its estimation relies on the consideration of alternate values of a parameter in a model of the experiment. To assess systematic uncertainty in the untagged analysis, ATLAS reports that they

> performed pseudo-experiments (PEs) with simulated samples which included the various sources of uncertainty. For example, for the JES uncertainty, PEs were performed with jet energies scaled up and down according to their uncertainties and the impact on the cross-section was evaluated. (ATLAS [2012d], p. 250)

Although a different methodology was used in the tagged analysis, that methodology likewise relied on variation of parameter values in a model of the experiment.

It is precisely the strategy of varying the inputs to a model-based estimation procedure within the bounds of what is possible, given the limitations on one's knowledge, that is indicative of a robustness analysis in this context.

Acknowledgements

I am grateful for helpful discussions with and suggestions from Giora Hon, Allan Franklin, Eran Tal, Luca Mari, Fabien Grégis, Wendy Parker, Robert Northcott, Anna Alexandrova, Chiara Lisciandra, Caterina Marchionni, Aki Lehtinen, Lorenzo Casini, Jonah Schupbach, and Jacob Stegenga. Conversations with Hugo Beauchemin have been especially valuable. This paper developed out of a primordial version presented at a workshop on modelling at the LHC in Wuppertal, Germany, held under the auspices of the Epistemology of the LHC project. Previous versions were presented at the September 2014 Workshop on Robustness Analysis in Helsinki, Finland, the Society for Philosophy of Science in Practice meeting in Aarhus, Denmark in June 2015, and the Informal Aspects of Uncertainty Assessment workshop in Cambridge, England in May 2016. Two anonymous referees for this journal provided very helpful feedback and suggestions.

Saint Louis University

Department of Philosophy

3800 Lindell Blvd.

St. Louis, MO 63108 U.S.A.

staleykw@gmail.com

References

ATLAS [2012a]: 'Measurement of the t-Channel Single Top-Quark Production Cross Section in pp Collisions at $\sqrt{s}$ = 7 TeV with the ATLAS Detector', *Physics Letters B*, **717**(4–5), pp. 330 – 350.

ATLAS [2012b]: 'Measurement of the Top Quark Pair Cross Section with ATLAS in pp Collisions at $\sqrt{s}$ = 7 TeV Using Final States with an Electron or a Muon and a Hadronically Decaying τ Lepton', *Physics Letters B*, **717**(1—3), pp. 89 –108.

ATLAS [2012c]: 'Measurement of the Top Quark Pair Production Cross Section in pp Collisions at $\sqrt{s}$ = 7 TeV in Dilepton Final States with ATLAS', *Physics Letters B*, **707**(5), pp. 459–477.

ATLAS [2012d]: 'Measurement of the Top Quark Pair Production Cross-Section with ATLAS in the Single Lepton Channel', *Physics Letters B*, **711**(3–4), pp. 244–263.

Barlow, R. [2002]: 'Systematic Errors: Facts and Fictions', Unpublished. <arXiv:hep-ex/0207026>

Beauchemin, P.-H. [2015]: 'Autopsy of Measurements with the ATLAS Detector at the LHC', *Synthese*, pp. 1–38.

Botje, M., Butterworth, J., Cooper-Sarkar, A., de Roeck, A. *et al.* [2011]: 'The PDF4LHC Working Group Interim Recommendations', Tech. rep. ArXiv:1101.0538.

Boumans, M. and Hon, G. [2014]: 'Introduction', in M. Boumans, G. Hon and A. Petersen (*eds*), *Error and Uncertainty in Scientific Practice*, London: Pickering and Chatto, pp. 1–12.

CMS [2011]: 'First Measurement of the Cross Section for Top-Quark Pair Production in Proton–Proton Collisions at $\sqrt{s} = 7$ TeV', *Physics Letters B*, **695**(5), pp. 424–443.

CMS [2013]: 'Measurement of the $t\bar{t}$ Production Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV with Lepton + Jets Final States', *Physics Letters B*, **720**(1–3), pp. 83–104.

Cousins, R. D. [1995]: 'Why Isn't Every Physicist a Bayesian?', *American Journal of Physics*, **63**, pp. 398–410.

Cousins, R. D. and Highland, V. L. [1992]: 'Incorporating Systematic Uncertainties into an Upper Limit', *Nuclear Instruments and Methods in Physics Research*, **A320**, pp. 331–335.

Cranmer, K. [2003]: 'Frequentist Hypothesis Testing with Background Uncertainty', in L. Lyons, R. Mount and R. Reitmeyer (*eds*), *Statistical Problems in Particle Physics, Astrophysics, and Cosmology: Proceedings of PHYSTAT 2003*, Stanford, CA: SLAC, pp. 261–264.

Egan, A. and Weatherson, B. (*eds*) [2011]: *Epistemic Modality*, New York: Oxford University Press.

James, F. and Roos, M. [1991]: 'Statistical Notes on the Problem of Experimental Observations Near an Unphysical Region', *Physical Review D*, **44**, pp. 299–301.

JCGM Working Group I [2008]: *Evaluation of Measurement Data – Guide to the Expression of Uncertainty in Measurement*, Joint Committee for Guides in Metrology. <http://www.bipm.org/en/publications/guides/gum.html>

JCGM Working Group II [2012]: *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms*, Joint Committee for Guides in Metrology. <http://www.bipm.org/en/publications/guides/vim.html>

Kratzer, A. [1977]: 'What 'Must' and 'Can' Must and Can Mean', *Linguistics and Philosophy*, **1**, pp. 337–55.

Mari, L. and Giordani, A. [2014]: 'Modelling Measurement: Error and Uncertainty', in M. Boumans, G. Hon and A. Petersen (*eds*), *Error and Uncertainty in Scientific Practice*, London: Pickering and Chatto, pp. 79–96.

Morrison, M. [2015]: *Reconstructing Reality: Models, Mathematics, and Simulations*, New York: Oxford University Press.

Parker, W. S. [2009]: 'Confirmation and Adequacy-for-Purpose in Climate Modelling', *Aristotelian Society Supplementary Volume*, **83**(1), pp. 233–249.

Parker, W. S. [2011]: 'When Climate Models Agree: The Significance of Robust Model Predictions', *Philosophy of Science*, **78**(4), pp. 579–600.

Parker, W. S. [2012]: 'Scientific Models and Adequacy-for-Purpose', *The Modern Schoolman*, **87**, pp. 285–293.

Parker, W. S. [2015]: 'Computer Simulation, Measurement, and Data Assimilation', *The British Journal for the Philosophy of Science*.

Rabinovich, S. [2007]: 'Towards a new edition of the "Guide to the expression of uncertainty in measurement"', *Accreditation and Quality Assurance*, **12**(11), pp. 603–608.

Sinervo, P. [2003]: 'Definition and Treatment of Systematic Uncertainties in High Energy Physics and Astrophysics', in L. Lyons, R. Mount and R. Reitmeyer (*eds*), *Statistical Problems in Particle Physics, Astrophysics, and Cosmology: Proceedings of PHYSTAT 2003*, Stanford, CA: SLAC, pp. 122–129.

Staley, K. W. [2004]: 'Robust Evidence and Secure Evidence Claims', *Philosophy of Science*, **71**, pp. 467–488.

Staley, K. W. [2012]: 'Strategies for Securing Evidence through Model Criticism', *European Journal for Philosophy of Science*, **2**, pp. 21–43 10.1007/s13194-011-0022-x.

Staley, K. W. [2014]: 'Experimental Knowledge in the Face of Theoretical Error', in M. Boumans, G. Hon and A. Petersen (*eds*), *Error and Uncertainty in Scientific Practice*, London: Pickering and Chatto, pp. 39–55.

Staley, K. W. and Cobb, A. [2011]: 'Internalist and Externalist Aspects of Justification in Scientific Inquiry', *Synthese*, **182**, pp. 475–492 10.1007/s11229-010-9754-y.

Tal, E. [2012]: *The Epistemology of Measurement: A Model-Based Account*, Ph.D. thesis, University of Toronto, Toronto.

Tal, E. [2016]: 'Making Time: A Study in the Epistemology of Measurement', *The British Journal for the Philosophy of Science*, **67**(1), pp. 297–335.

Vezér, M. A. [2016]: 'Computer Models and the Evidence of Anthropogenic Climate Change: An Epistemology of Variety-of-Evidence Inferences and Robustness Analysis', *Studies in History and Philosophy of Science Part A*, **56**, pp. 95–102.

Willink, R. [2013]: *Measurement Uncertainty and Probability*, New York: Cambridge University Press.