

Why Trust a Simulation?

Models, Parameters, and Robustness

in Simulation-Infected Experiments

Florian J. Boge

Abstract

Computer simulations are nowadays often directly involved in the generation of experimental results. Given this dependency of experiments on computer simulations, that of simulations on models, and that of the models on free parameters, how do researchers establish trust in their experimental results? Using high-energy physics (HEP) as a case study, I will identify three different types of robustness that I call conceptual, methodological, and parametric robustness, and show how they can sanction this trust. However, as I will also show, simulation models in HEP themselves fail to exhibit a type of robustness I call inverse parametric robustness. This combination of robustness and failures thereof is best understood by distinguishing different epistemic capacities of simulations and different senses of trust: Trusting simulations in their capacity to facilitate credible experimental results can mean accepting them as means for generating belief in these results, while this need not imply believing the models themselves in their capacity to represent an underlying reality.

- 1 *Introduction*
- 2 *Robustness Analysis: Taking Stock and Looking Ahead*
- 3 *Conceptual Robustness*
 - 3.1 *The modelling challenge*
 - 3.2 *Never trust a (single) generator!*
- 4 *Methodological Robustness*
 - 4.1 *Convergence of results*
 - 4.2 *Discrimination*
 - 4.3 *Security versus informativeness*
- 5 *Robustness and Parameters*
 - 5.1 *Universal or not?*
 - 5.2 *Tuning for accuracy*

6 Trust and Its Limits

6.1 Commonality, confirmation, and two senses of trust

6.2 Instrumental reliability

6.3 Assumption convergence: A rejoinder?

7 Conclusions

1 Introduction

Computer simulations (CSs) are nowadays often directly involved in the generation of experimental results. This fact has recently been recognized by philosophers for raising a number of epistemological challenges: Simulated data are often treated as equivalent to experimental data, and this requires an analysis as to the appropriateness of such treatments (Morrison [2015]; Massimi and Bhimji [2015]; Parker [2017]). Devising a CS presupposes a whole chain of models that includes manifold successive modifications, from simplifications and approximations to outright artificial constructions (Winsberg [1999], [2010]; Lenhard [2007]; Morrison [2009]; Boge and Zeitnitz [forthcoming]), and this introduces sources of potential error from incorrect modelling to malfunctioning hard- and software. Finally, (virtually) every CS makes use of parameters that have to be fixed empirically, and this raises the question of how these parameters influence the epistemic status of corresponding experimental results, and of the models used to facilitate them (Hasse and Lenhard [2017]; Gueguen [2019]; Ritson and Staley [2020]).

So given this dependency of experiments on CSs, the dependency of CSs on modelling assumptions, and the dependency of the models on parameter choices, how do researchers establish trust in their experimental results and what does this tell us about the models used? I will here focus on robustness analysis (RA), characterized as a ‘method for determining which models make trustworthy predictions’ by Weisberg ([2006], p. 731), as a key to understanding the prospects and limitations of trusting CSs in the generation of simulation-infected experimental results.

Like Staley ([2020]), I will use experimental high-energy physics (HEP), the study of interactions between material particles in high-energy collisions, as a case study. For Morrison ([2015], p. 287) displays CSs as a *conditio sine qua non* therein: They are used to estimate

the discovery potential of detector components before these are built, define the ‘background’ data to be expected in the absence of phenomena of interest for statistical analyses, or even what signatures of interest might look like. Thus, ‘simulation plays a significant role *in* the experimental setting’ (Morrison [2015], p. 288).

I will argue for a differentiated view, according to which (HEP) researchers justifiably trust their experimental results without having to invest outright belief in the simulation models used to facilitate them. I hence distinguish two senses of trust that could be invested in simulation models: trust in their capacity to facilitate credible experimental results, and trust in their capacity to faithfully represent an underlying reality. That these can fall apart in principle should be evident: an experimenter’s skill, say, or a measuring instrument, can be trustworthy in the former sense without even being suitable for the latter. However, simulation models are generally intended as (more or less accurate) representations; so it is a non-trivial observation if they turn out trustworthy in the former but not in the latter sense.

To carefully distinguish between both capacities, I will identify four types of robustness, three of which concern experimental results: (i) their insensitivity to varying modelling assumptions in the definition of simulation programmes used in their generation; (ii) to different uses of CS therein; and (iii) to changes in the values of the models’ free parameters.¹ The fourth, however, concerns the parameters themselves, namely, (iv) the insensitivity of their values to varying experimental conditions in which the models are used.

Below, I will call (i) conceptual robustness (CR), as it concerns the conceptualization of simulation targets; (ii) methodological robustness (MR), as it concerns simulation-based methods for extracting experimental information; (iii) parametric robustness (PR), for obvious reasons; and (iv) inverse parametric robustness (IPR), as it concerns the parameters, not the results gained with them.

CR, MR, and (some sense of) PR establish a rather solid basis for HEP researchers to trust their experimental results. In contrast, IPR is doubtful, and this has implications for the models’ epistemic status. As I shall argue, the relevant simulation models are best understood as

¹ (i) and (iii) are closely related to what Weisberg ([2013], pp. 159–61) calls parameter and structural robustness, the salient difference being that I am here interested in simulation-infected experimental results, not model behaviour alone.

reliable cognitive instruments, not faithful representations of the goings on in the ‘beam pipes’ of colliders.

The structure is as follows: After a recap of some relevant results from the debate on RA (Section 2), the first, ‘analytic’ part (Sections 3–5) establishes the four types of robustness and the sense of justification they convey. The second, ‘critical’ part (Section 6) then inquires about the epistemic status of the models used in relevant inferences. I also offer an outlook, in the conclusion, on why the results are not only interesting with respect to the article’s main case study but to simulation-heavy fields more generally.

2 Robustness Analysis: Taking Stock and Looking Ahead

Following Wimsatt ([1981]), the ‘robustness’ of an experimental or theoretical result means its invariance under changing means of detection. Robustness analysis is the process of ‘exploring the differences in past means of detection that have had no (or a negligible) effect on the result—and also [...] exploring variations that have made a difference to the result’ (Schupbach [2016], p. 277). RA has been applied to models (for example, Lloyd [2009], [2010]; Weisberg [2006]; Weisberg and Reisman [2008]; Parker [2011]), experiments (Schupbach [2016]; Karaca [2022]; Staley [2004b], [2020]), and CSs (Muldoon [2007]; Durán and Formanek [2018]; Gueguen [2019]; Lehtinen and Kuorikoski [2007]). In HEP, these issues become entangled, and subjecting models and simulation to RA can in principle serve the goal of inferring something about experimental results or the status of simulation models.

Famously, there has been considerable debate about what an RA across models can establish. In response to Levins ([1966]), Orzack and Sober ([1993]) had objected that Levins’s discussion made RA look like a method of non-empirical confirmation—something not easily acceptable to most philosophers. Furthermore, there is a long-standing debate on how to assess the diversity of different means detection involved in RA, as well as their independence (see Schupbach [2016], Sections 2.1–2.4).

These concerns have been forcefully answered by Schupbach ([2016]): If a certain result is invariant under changing means of detection, then this successively rules out different explanations of that result in terms of properties peculiar to any given detection. Competing

explanations of the result's occurrence thus need not be fully independent, so long as accepting one rules out the other *qua* explanation.² Furthermore, models and empirical procedures can in principle be treated on a par as means of detection, whence the consistency between measurement results and model predictions can contribute to the confirmation of a result by safeguarding against reliance on either modelling or measurement errors (Winsberg [2018]). However, this (strong) sense of confirmation is obviously not non-empirical.

Most experimental results are quantitative, and two means of detection never yield the exact same values. Hence, I prefer talk of 'insensitivity' rather than invariance, meaning that individual results, subjectable to RA, do not so strongly depend on the individual means of detection that they become statistically inconsistent with one another.³ Salient experimental results in HEP are evidence for, or observations of, new particles, measurements of, say, anomalous production rates for certain final-state particles, or bounds established on certain quantities of interest. 'Evidence' and 'observation' are technical terms in this context, meaning a discovery significance above 2.8 or 5 standard deviations, respectively (Franklin [2013]). Since observations and evidence thus rely on measured excesses of events in specific regions of certain kinematic variables, I will treat all these results as measurements. Furthermore, 'measurements' are rarely ever direct comparisons between physical quantities, but usually involve complex modelling and theoretical inferences (Morrison [2009], [2015]; Parker [2017]; Tal [2012], [2016]).

In a recent article, Staley ([2020]) has argued that such results gain their credibility in part from being associated with a measure of systematic uncertainty, which can be fruitfully viewed as the result of a RA, reflecting the impact of choices in the modelling assumptions that facilitate the inference to the desired result.⁴ This is consistent with Karaca's (Karaca [2022], p. 7) verdict that robustness 'comes in degrees in the sense that its extent indicates the extent of convergence of [...] results obtained through different means of detection'. Hence, a result's systematic uncertainty may be taken to indicate this degree of convergence, when the different

² See, however, (Stegenga and Menon [2017]) for an alternative in terms of conditional probabilistic independence.

³ This will become clearer in Section 4.1.

⁴ In contrast, statistical uncertainty is 'a characteristic of the distribution of outputs of a single model of a measurement process' (Staley [2020], p. 92).

means are different models facilitating the experimental inference.

Staley ([2020]) presupposes what he calls the 'secure evidence framework', according to which an epistemic agent needs to evaluate claim C that data \mathbf{x} count as evidence for hypothesis H in relation to her epistemic situation K and an induced set of epistemically possible scenarios Ω . Defining the result's systematic and statistical uncertainty is key to weakening C , which, equivalently, means rendering it independent from potentially erroneous assumptions in Ω . For the narrower context of generating a quantitative experimental result (QER), this means that if a range of different models of an experiment that yields data \mathbf{x} all give rise to results within an interval $[\lambda - \delta, \lambda + \delta]$, then stating the result as quantity Q taking on value $Q = \lambda \pm \delta$ is secure from being dependent on any errors contained in these models.

Without following suit with all the details of Staley's framework, I will here embrace the notion that the models used to infer a QER reflect what is initially considered epistemically possible, and that RA can serve the purpose of determining the QER's systematic uncertainty. However, if $[\lambda - \delta, \lambda + \delta]$ exhausts a significant part of the real line, then it is hardly sensible to speak of the 'convergence of results obtained through different means of detection'. According to Staley ([2014], pp. 48–54, [2020], p. 103), the strengthening of the premises of an experimental inference, by carefully ruling out models whose inclusion would lead to a wider spread $\pm\delta$, is key to safeguarding against this. I believe that a combination of RAs can also do the trick. I will take up this issue, which has a close connection to the notion of measurement accuracy, in Section 4.3.

3 Conceptual Robustness

3.1 The modelling challenge

RA with respect to models and their trustworthiness has traditionally been employed in the context of complex sciences without an encompassing fundamental theory, where different models necessarily rely on strongly simplifying and idealizing assumptions. Examples include ecology (Weisberg [2006]; Weisberg and Reisman [2008]), climate science (Lloyd [2009], [2010]; Parker [2011]; Winsberg [2018], [forthcoming]), or economy (Kuorikoski *et al.* [2010]; Lehtinen and Kuorikoski [2007]). However, with the electroweak theory and quantum chromody-

namics (QCD), jointly called the ‘standard model’ (SM), HEP features a rather solid theoretical basis. So whence the need for RA?

In fact, the vastness of decays, the nature of the strong coupling (which allows theoretical computations only for very high energies), and the complexity of the instrumentation required to tackle all these decays together necessitate various models that are ‘phenomenological’ in the sense of not being strongly bound to theory in their construction (Suárez and Cartwright [2008]; Karaca [2013]). CSs in HEP comprise two general kinds of simulation based on these models: Monte Carlo event generators (short: ‘generators’), simulating the physics inside the ‘beam pipe’ (the metallic tube in which particles collide); and detector simulations, simulating the interaction of sufficiently stable decay products with detector components (ATLAS Collaboration [2010]).

Generators are further divided into several stages, called hard scatter, underlying event, parton shower, and hadronization. These are heuristically assumed to separate, but feature manifold connections that ultimately impose the need for mutual recalibration (Boge and Zeitnitz [forthcoming]). The ‘hard scatter’ is the elementary scattering process with the highest momentum transfer; an interaction between two ‘partonic’ constituents (quarks or gluons) of the scattering protons. This is usually the process of interest, because here one may expect the production of sought-for particles. The ‘underlying event’ means additional, less energetic scatterings, or even particles remaining from previous or differently located scatterings. The ‘parton shower’ is the emission of additional gluons by the accelerating and decelerating coloured particles before and after the primary interaction. This may be understood in analogy to the ‘Bremsstrahlung’ of accelerating and decelerating electrical charges, but with the radiated particles (gluons) being coloured and emitting further radiation. Finally, at lower energies, the confining effects of QCD become important, and partons begin to hadronize, that is, combine into complex particles, most of which subsequently decay into more stable ones.

All this goes on before any interaction with the detector. The LHC detectors are giant apparatuses comprising various technologies. Accordingly, detector simulation takes inputs also from data, atomic, nuclear, and solid state physics (see GEANT4 Collaboration [2017]).

Basically all models in this context require extra-theoretic considerations. The hard pro-

cess builds on perturbative QCD calculations to some fixed order in a series-development of elementary cross-sections. However, proton–proton cross-sections are thereby assumed to ‘factorize’ into weighted sums of parton-level cross-sections; an assumption actually only proven for a small range of cases (Schwartz [2014], p. 685; Collins [2011], pp. 1–2). The resulting probability weights (called ‘parton density functions’), furthermore, can only be obtained by fitting parametric functions to a broad range of data (also Staley [2020]; Ritson and Staley [2020]). Similarly, shower models utilize an approximate QCD result as the basis of a Markov chain in which partons are radiated at distinct times (see Höche [2016], p. 242). But the resulting probability functions diverge in the limit where the approximation becomes exact, and CSs include a pragmatic cut-off, corresponding to the resolution of the detector.

Theoretical results are especially sparse for hadronization,⁵ since perturbative QCD breaks down at lower energies. Furthermore, these models ‘inextricably couple’ (Dissertori *et al.* [2003], p. 172) to other models in the chain, which forestalls validating them independently. Hence the need for RA. However, for an RA to even possibly succeed, different models of the same process need build on different idealizing and simplifying assumptions, so that an RA can exclude that a common result is an artefact of any particular set of such assumptions. Focusing on hadronization as the weakest link in the simulation chain, and ‘string’ and ‘cluster’ models as the currently most common ones, I shall briefly expand on this.

The main differences here reside in the treatment of the colour field. String models focus on a particular limiting behaviour of this field and largely ignore its quantum nature. They model it as a flexible string, binding pair-produced quarks, with dynamics determined by a parametric ‘fragmentation function’ that specifies under which conditions the string tears. This event indicates the creation of two (unstable) mesons, corresponding to the string fragments.

In contrast, cluster models focus on certain quantum aspects, namely, the prediction of colour entanglement between neighbouring quarks, also called ‘preconfinement’ into ‘clusters’, from a QCD approximation that formally takes the number of colour charges to infinity. The resulting clusters can be shown to follow a universal (not decay-specific) mass distribution, which the

⁵ This is similarly true of the underlying event and detector models; but the modelling freedom in the former is constrained by the similarity to the hard scatter, and the latter can in part be validated directly (Boge and Zeitnitz [forthcoming], pp. 28–29).

simulation samples. Thus, cluster models completely neglect the field dynamics.

Similarly, whereas cluster models split gluons from the shower stage non-perturbatively into quark–anti-quark pairs, and so essentially ignore them, string models at least preserve them as an aspect of the semi-classical, string-like field: as ‘kinks’ (or transverse excitations) in the string, which leads to the generic prediction of a more frequent occurrence of hadrons between quark and gluon.

3.2 Never trust a (single) generator!

Given these preliminaries, the case for CR is straightforward, and has in part been covered by Staley ([2020], p. 110). Physicists have various generators at their disposal, and comparing the results gathered from using different generators is basically mandatory:

Monte Carlo event generators are complicated programs that will almost inevitably contain bugs, incorrect assumptions and ill-chosen parameters. It is therefore vital that a user does not take any results at face value. As a minimum at least two completely independent programs should be used in any physics study. (Dissertori *et al.* [2003], p. 204; emphasis omitted)

A number of things are worth emphasizing though. First, the mention of ‘completely independent programs’ is well in line with Levins’s famous slogan that the ‘truth’ established by RA is ‘the intersection of independent lies’ (Levins [1966], p. 423). However, there are various connections between different hadronization models and programmes, and so talk of complete independence is far-fetched. As explained in Section 2, this need not pose a problem, so long as RA across different models can rule out the explanation that a certain result arises as an artefact of the idealizing and simplifying properties of these models.

Second, establishing this sort of CR is piecemeal work, and an RA across models needs to be performed for every data–model combination. A recent study on the top quark-mass, for example, employed a newer set of ‘special-purpose’ generators, capable of computing next-to-leading order matrix elements. But the results varied grossly when these were interfaced with two different ‘general-purpose’ generators—‘to the point of drastically altering the conclusions’ (Ravasio *et al.* [2018], p. 3). Hence, depending on the data–model combination, CR may fail.

Third, the differences in simulation results, due especially to the limited understanding of the hadronization phase, are generally used to define a ‘hadronization uncertainty’ (Dissertori *et al.* [2003], p. 244; emphasis omitted), in perfect agreement with Staley ([2020])’s assessment. Similarly, ‘in order to assess a meaningful theoretical error [...], the use of different shower models [...] is mandatory’ (Ravasio *et al.* [2018], p. 4).

However, this is considered a merely ‘pragmatic approach to assessing the uncertainties associated with hadronization’ (Dissertori *et al.* [2003], p. 172), or simulation more generally. There is, in other words, no solid *a priori* reason for determining uncertainties in this or any other way. In addition, one frequently finds that the prescription uttered by Dissertori *et al.* is obeyed only in a minimal sense, that is, that only two sets of generators are used for robustness checks. Besides such concerns as computational expensiveness, this reliance on a minimal number of different models may be rooted in the vastness of the differences between them.⁶

Concluding that a common result is unlikely to arise coincidentally from two very different models is well in line with the general thinking underpinning RA (for example, Kuorikoski *et al.* [2010], p. 544), albeit with the unlikely coincidental nature of the result rooted in the vastness of differences rather than the number of models. Still, it is not clear just how safe an experimental conclusion relying on a limited number of models like this really is.

4 Methodological Robustness

4.1 Convergence of results

Given the aforementioned restrictions in relying on CR alone, it is understandable that this is not the only sort of RA physicists in HEP perform. To see this, consider the different measurements of the top quark mass, m_t , displayed in Figure 1. Neither do the central values of m_t coincide, nor do all margins of total uncertainty overlap. But most of them do, and they all overlap with the uncertainty of the combined world average. All these different measurements are hence consistent with the central claim that $m_t = 173.34 \pm 0.76$ GeV.

To infer these different values, ATLAS and CMS used several decay channels, as specified

⁶ For example, HERWIG not only uses cluster rather than string hadronization, but also an angular ordering in the parton shower rather than the momentum ordering of PYTHIA (Bähr *et al.* [2008]; Sjöstrand *et al.* [2008]). This means slicing the radiation process in different ways, and is but one of many theoretically allowed choices.

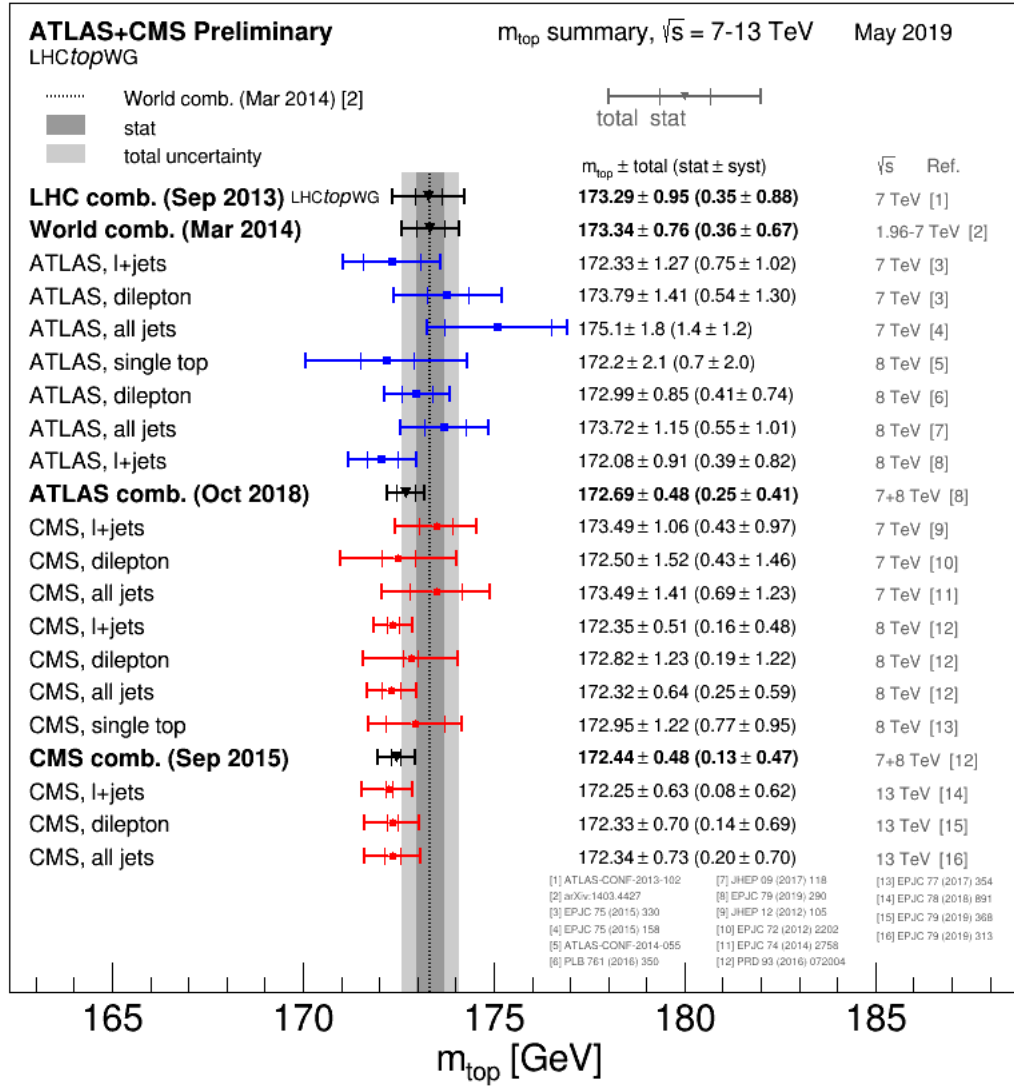


Figure 1. Summary of measurements of m_t by the LHC top working group (CERN [2019]). Colour figure available online

by comments like ‘ ℓ (epton) + jets’ or ‘single top’. This means that next to the fact that several generators are used in each individual measurement, there are also different processes simulated each time. Furthermore, the standard tool for simulating detector effects, the GEANT simulation toolkit, comprises a vast range of detailed physics models (see GEANT4 Collaboration [2017]), and interactions vary with decay chains—as do materials and geometry between ATLAS and CMS (see Seiden [2012]). Finally, the combined world average also contains results from the D0 and CDF collaborations at Fermilab, which used protons and anti-protons, scattered at much lower energies, with final-state products measured by yet other detectors (see ATLAS, CDF, CMS and D0 Collaborations [unpublished]).

Additionally, another sort of variation is performed across several measurements: Two stand-

ard analyses are the template and matrix element method, and the roles played by simulated data therein differ crucially. In the first, a range of input values for m_t are used to generate simulated histograms. These are then used as ‘templates’ for what actual data should look like, given a value of m_t , and one infers the best value by means of likelihood maximization (see ATLAS Collaboration [2012]). Hence, simulated data are used in a direct, positive fashion in the template method.

In the matrix element method, one instead calculates from the theory as far as possible, using different values for m_t , and predicts the states to be measured after interaction with the detector from a parametric ‘response function’. This function also subsumes shower and hadronization effects, thus by-passing a direct involvement of relevant models (Fiedler *et al.* [2010], p. 204–5). The parameters of that function are, however, usually fixed using simulated data (D0 Collaboration [2006]), and so CSs still enter indirectly. They also enter negatively in defining the distribution of events to be expected on the ‘background only’ hypothesis (Fiedler *et al.* [2010], p. 212).

4.2 Discrimination

Given all these differences, the consistency between the different results is certainly remarkable. However, by itself, this does not yet rule out that the convergence is due to dependencies between the different usages of CSs. It seems clear that neither usage of CSs is fully independent of the other. For instance, they do not seem to satisfy full ‘ontic independence’ (Stegenga and Menon [2017], p. 414), that is, that ‘the multiple lines of evidence depend on different materials, assumptions, or theories’. But are they ‘independent enough’ to ensure that MR is not a kind of ‘pseudo-robustness’ (Karaca [2022], p. 4; Staley [2004b], p. 473)?

In order to safeguard against pseudo-robustness, different measurements, making use of CSs in different ways, should not be ‘correlated with each other to the extent that they yield convergent results when they should not’ (Karaca [2022], p. 2). Both Staley ([2004b]) and Karaca ([2022]) consider the case of different b -tagging algorithms, which were compared as to their ‘mistags’ in the top-quark discovery. Since mistags differed among different algorithms, the algorithms were not simply ‘correlated all across the board’. However, in establishing MR, CSs

are used to simulate different things, whence this sort of ‘discriminant validation’ (Campbell and Fiske [1959], p. 81) is not feasible.

Another, encompassing sense of discrimination is captured by Schupbach ([2016], pp. 291–06), who considers several detections R_1, R_2, \dots, R_{n-1} , so far compatible with hypotheses H and H' , where H' signifies the hoped-for explanation of the apparently robust result, and H is any rival explanation. H and H' could in principle logically coexist, but accepting H' implies that H becomes a bad explanation for the past detections. Discrimination now means that relative to the remaining $n - 1$ detections, detection R_n can really only be explained by H' , not H . This is an iterative process: each of the n detections may rule out a different hypothesis H , so as to only leave H' as a plausible explanation.

For robustness to be more than just apparent, different detections themselves hence need to discriminate against different hypotheses H , that is, against explanations of the convergent result in terms of artefacts introduced by either detection. So to rule out, say, that the result is purely due to b -tagging algorithms’ spurious general agreement, one could instead also check these measurements’ consistency with measurements without b -tagging.

In the case of the different usages of CSs, the different hypotheses H ruled out are that the measurement result is influenced by the given CS to the extent of becoming an artefact. In contrast, the hypothesis H' to be inferred in case of success is the central result that is consistent with all the individual measurements (here: $m_t = 173.34 \pm 0.76$ GeV). H and H' are logically compatible: The actual value might simply happen to lie in the inferred interval even if individual measurements predict this on the basis of simulation-related artefacts.

That the different detections of m_t are indeed relevantly discriminating can be seen as follows: Simulating only the background, as in the matrix element method, means eliminating the possibility that the value $m_t = 173.34 \pm 0.76$ GeV is inferred purely due to a misidentification of the signal shape on the basis of modelling errors (as might happen in the template method). Similarly, inferring values consistent with H' both from templates that contain a single lepton and templates that contain only hadronic jets means ruling out that the result is inferred purely due to a misidentification of the properties of the lepton. The ℓ +jets and all jets channels also do not depend equally strongly on the choice of hadronization model, and the corresponding

uncertainties differ by 350 MeV (ATLAS, CDF, CMS and D0 Collaborations [unpublished], p. 20). Hence, as these examples demonstrate, the aforementioned differences render pseudo-robustness at least very improbable.

4.3 Security versus informativeness

Recall now how a quantitative RA can only succeed if results from different detections lie in sufficient proximity to one another. In close connection to this observation, Parker ([2017], p. 279) points out that ‘a well-motivated uncertainty estimate [...] [also] indicates the degree to which the measuring process is expected to be informative’, where by ‘informativeness’, she roughly means the extent to which a result narrows down the epistemic background possibilities an agent can coherently entertain. Similarly, Staley ([2020], pp. 105–6) concerns himself with a measurement’s sensitivity to the phenomena of interest, construed as ‘the ability of a measurement result to inform our answers to substantive physics questions’.⁷

Thus, there appear to be two distinct pay-offs to a quantitative RA: (i) securing the result from being severely wrong by quantifying the impact of possible errors in its generation; and (ii) establishing its informativeness by exhibiting the extent to which individual detections converge. However, (i) and (ii) pull in opposite directions: a broader uncertainty margin is ‘safer’, as it admits to more potential errors; but it arguably renders the result less informative.

Another way of phrasing things is in terms of accuracy. Following Tal ([2012], pp. 30–32), there are several different readings of ‘measurement accuracy’, and construed epistemically, ‘accuracy’ means that the design and use of an instrument or procedure ‘warrant the attribution of a narrow range of [...] values to objects’ under study. This is not to be conflated with precision, which is just one aspect of epistemic accuracy, with imprecision being ‘caused by uncontrolled variations to the equipment, operation or environment when measurements are repeated’. In contrast, uncertainty, construed as an antonym to epistemic accuracy, can obviously also have systematic sources. Furthermore, ‘accuracy’ can also be meant in a comparative sense, that is, as a result’s reproducibility ‘under controlled variations to measurement

⁷ Following (Beauchemin [2017]), he means by ‘sensitivity’ that the result’s uncertainty is small enough to discriminate between different theoretical predictions (see Staley [2020], p. 106). It is easily seen that this closely corresponds to the notion of informativeness I invoke below.

conditions and methods’.

However, epistemic accuracy thus also delivers a (though not ‘the’) measure of a QER’s informativeness: A QER with small uncertainty is not only epistemically accurate, but also informative about the phenomena of interest, since it rules out a broader spectrum of models or theories. If a QER yields $\hat{Q} \in \Delta$, but theory T predicts that quantity Q actually has a value in Γ , where $\Gamma \cap \Delta = \emptyset$, T may stand refuted unless one finds reasons to broaden the uncertainty margin (say, by reassessing the auxiliary hypotheses needed to derive $Q \in \Gamma$ from T).

As mentioned before, Staley ([2020], Section 7) addresses these issues in terms of ‘strengthening strategies’ that (mostly) consist in ruling out possibly erroneous premises empirically, thereby decreasing the uncertainty margin. However, this too is not generally feasible for HEP simulation models, due to their ‘inextricable coupling’. Luckily, physicists have another ace up their sleeves. To see this, consider m_t ’s combined world average again, which was computed by the ATLAS, CDF, CMS and D0 Collaborations ([unpublished]) using an algorithm that approximates the best linear unbiased estimate (‘BLUE’); that linear combination of the different measured values that minimizes overall uncertainty. Hence, the resulting uncertainty measure specifies the minimum uncertainty that has to be assumed when all measurements are taken into account.

This uncertainty is larger than that of the overall ATLAS and CMS values, but smaller than that of their combined value. It is also smaller than that of more than 80% of the results from individual decay chains. Hence, by replacing the epistemic accuracies associated with the individual results by a comparative measure of accuracy that establishes their consistency to within a rather narrow span, a combined result is generated that is not only more secure but also more informative than most individual results.⁸ In other words: by directly determining the extent of convergence between individual results, physicists achieve a balance between security and informativeness from a skilful combination of RAs, without the need to appeal to external strengthening strategies.

⁸ I admit that the increase in informativeness is accompanied by a decrease in security at the level of meta-analysis: in using the BLUE method, certain non-trivial assumptions have to be made, such as results being Gaussian-distributed. However, the general epistemic intuition, that the best value must lie somewhere in between and should be associated with the minimum uncertainty necessitated by all results, would certainly be served by many other methods.

5 Robustness and Parameters

5.1 Universal or not?

CSs in HEP depend on a whole number of free parameters, and such an occurrence of parameters is ubiquitous in simulation-intensive sciences, but ‘has received surprisingly little attention from philosophers’ (Hasse and Lenhard [2017], p. 95). Free parameters bear the danger of spoiling predictivity: The success that string models can claim by predicting a specific distribution of hadrons from kinks in the string is arguably undermined by the fact that these predictions only obtain a quantitatively precise form once the model’s parameters have been fitted to data. Given that successful predictions are the hallmark of confirmation, this means trouble for parametrized models.

Of course, parameters are actually ubiquitous in science overall: With neutrino masses, the SM itself has twenty-seven free parameters, but this has not hindered its being hailed the most successful scientific theory to date. Three factors have arguably facilitated this immunity to the parameters’ success-undermining power: (a) they are physically meaningful quantities (such as masses and couplings); (b) there are certain dependencies between them that constrain some of them, once others have been fixed (they exhibit an exploitable coherence); and, most importantly, (c) ‘the Standard Model is an overconstrained system—we can test it by making enough measurements with enough precision’ (Schwartz [2014], p. 641).

When literally thought of as parameters of a model, (c) expresses the property I have called inverse parametric robustness (IPR) in Section 1: there is a set of values that can be consistently inferred from a range of rather different experimental settings. However, on account of (a), these ‘parameters’ of the SM are simply experimentally robust quantities in the traditional sense (Staley [2004b]; Schupbach [2016]; Karaca [2022]).

Now, are (a)–(c) also true of the parameters that occur in HEP simulation models? Let us address this step by step. First of all, one of the two fundamental string parameters was initially used to model a ‘string tension’, or rather, the uniform energy-density of the narrow flux tube into which the colour field is supposed to collapse on account of the model. This idea gained momentum from an analogy to an approximate result from quantum electrodynamics, but the

approximation is known to be poor, and predictions based on the physically suggested value for the parameter are often off by large percentages (Dissertori *et al.* [2003], p. 166).

This type of process is characteristic of HEP simulation parameters. So, by and large, the situation appears to be as in thermodynamics, where certain ‘models have parameters which describe physical quantities in the simple context for which they were initially developed [...] But in the complex context in which they are used as fitting parameters, the original physical meaning is lost’ (Hasse and Lenhard [2017], p. 109).

So much for (a). What about (b), the parameters’ exploitable coherence? As emphasized in Section 3.1, HEP simulation parameters are ‘closely entangled’ (Corcella *et al.* [2018], p. 489), so it is in general not feasible to fix some of them first and then use these to constrain others. Hence, neither (a) nor (b) seem to hold for HEP simulation parameters.

It could be argued, though, that this is not so bad, so long as parameters can at least be fixed once and for all: (b) seems mostly like a convenience, and the failure of (a) mostly says that these parameters do not have a fundamental meaning. But they might still offer a phenomenological (that is: coarse, non-fundamental) representation of the physics inside particle colliders.

In this connection, it is important to note the generic split between event generators and detector simulations. While it is in principle conceivable that the detector exerts a non-negligible influence, the modelling proceeds as if the physics inside the beam pipe fully decouples from what goes on inside the detector. Hence, to establish a phenomenological pendant of (a), it would at least be necessary to establish a uniquely best set of parameters across the different experiments at the LHC, where scattered particles and energies are identical.

In fact, there are even stronger claims to the universality of certain parameter values in the physics literature:

[...] one finds in practice that [...] a single set of parameters describes the data at a wide range of energies and processes [...] Thus, hadronization models tuned to e^+e^- annihilation, and lower energy hadron collider data, are highly predictive for LHC events. (Seymour and Marx [2015], pp. 309–10)

The ‘in practice’ and the ‘wide range of energies and processes’ here indicate that insofar as universality can be established, it is partially grounded in RA. More precisely, parameters are

initially ‘tuned’ or calibrated to certain reference data that can be understood without the simulation. This can be done in a variety of ways, for instance by manually adjusting them in direct comparison with data, or by randomly sampling from a large number of choices until an optimum value is reached, as determined by a goodness-of-fit metric. More systematic approaches, such as the PROFESSOR toolkit (Buckley *et al.* [2010]), optimize polynomial expansions of the Monte Carlo predictions in several orders of their parameters across large numbers of bin values for various observables.

Because of the appeal to reference data, ‘tuning’ means establishing a given choice of values experimentally, but in a way that might still depend on the peculiarities of the chosen data. The established parameter set is hence usually validated on further data. This is done, for instance, using the RIVET toolkit (Bierlich *et al.* [2020]), ‘a library of experimental analyses and of tools for calculating physical observables from an event record’ (Buckley *et al.* [2010], p. 332). Since each validation means testing whether the parameter set describes the relevant data reasonably well, a set of values that achieves a good fit across these varying experimental conditions can be called ‘robust’ in the sense of IPR.

Now I only said that claims to universality are partially grounded in an experimental RA, and the reason is that it is also theory-supported:

One can show that this is a consequence of preconfinement: the perturbative production and evolution of partons takes care of the process and energy-dependence and the transition from partons to hadrons is a local process, independent of these factors. (Seymour and Marx [2015], p. 310)

However, recall that preconfinement is only an approximate result, derived by taking the number of colour-charges to infinity. There are, in fact, also theoretical reasons for doubting IPR:

[...] it is not absolutely guaranteed that models which reproduce e^+e^- data work equally well in a colored environment like $t\bar{t}$ events at the LHC, where initial-state radiation, color reconnection and underlying event play a role. (Corcella [2019], p. 71)

This is not an academic point: The ‘Monash tuning’ for PYTHIA (Skands *et al.* [2014]), for instance, which mostly exploited LHC data rather than data from electron-positron scattering, led to significant changes in certain string parameters (with some parameters being more than doubled; see Skands *et al.* [2014], Table 4), and simultaneously to improvements in the correspondence with several data. Skands *et al.* ([2014], p. 1) also mention some ‘interesting

discrepancies' though, as well as several generic reasons for further improving the PYTHIA code, which 'should in principle be accompanied by a reevaluation of the model constraints' (Skands *et al.* [2014], p. 27). Accordingly, the tune is 'By no means [...] the final word', and the study gives a bold outlook on the difficulties associated with IPR: 'If [...] some data sets result in significantly different tune parameters, one has a powerful indication that the universality of the underlying modeling is breaking down [...] (Skands *et al.* [2014], p. 31)'. Still, differences between the Monash tune and the former default are accompanied by important physical differences: In the default tune, the scatterers were leptonic (and so not coloured), whence, for instance, no initial-state gluon radiation is to be expected. Hence, the above verdict could still be upheld at this point: so long as there was a uniquely best tune across different LHC experiments, where the phenomenology is the same, these parameters could have physical meaning.

This is far from true though: Each LHC collaboration, like ATLAS or CMS, has a whole range of different tunes for the same generator, depending on the details of use. An impressive—though far from exhaustive—overview can be gathered from websites and documentation files for different generators.⁹

To get an impression of the purpose dependency, it is instructive to consider four tunes devised by the ATLAS Collaboration ([2014]) to match underlying event data. When combined with one out of four parton density functions, specified in the study, each of these achieved a good fit to reference data. But they vary amongst each other by up to some 24% of the lowest value of some parameters, and there is also a mention of 'Additional tunes [...] in preparation for use with higher-order matrix elements' ([2014], p. 1).

Given the relatively big differences in these values, even for one specific purpose, the possibility of associating a definite phenomenological meaning to these parameters seems pretty much spoiled. There is, in other words, no solid ground for claiming IPR to hold across the LHC.

⁹ See, for instance, <home.thep.lu.se/~torbjorn/pythia82html/Tunes.html>.

5.2 Tuning for accuracy

The apparent failure of IPR clearly impacts the models' representational capacity: If their parameters were to describe physical properties of processes like hadronization or the parton shower, their values should not be sensitive to the detector used, or to the specific type of event simulated. Even more so, there should not be several different, but equally suitable, values for the exact same simulated event. However, does this failure of IPR also spoil the possibility of trusting these models in their QER-facilitating capacity?

To address this question, observe, for instance, that the greatest differences effected in measurable quantities by the Monash tuning were in the order of 10% (Skands *et al.* [2014], p. 29). This is certainly not nothing, but also a big stretch from doubling, or even 24%. Furthermore, large uncertainties in parameters certainly can influence experimental results: In the 1990s, Tevatron data were interpreted as indicating evidence for a new level of sub-quark particles—which evidence vanished when certain parton density functions were recalibrated on additional data (Dorigo [2016], Chapter 9). However, uncertainty propagation can result in only small contributions to quantities of interest, which would safeguard against such an effect. Hence, even if IPR fails, isn't PR, the insensitivity of results to changes in parameter values, enough to render at least simulation-infected experimental results trustworthy?

I believe this is ultimately correct, but given the failure of IPR, PR itself enjoys a subtle status in HEP. To see this in detail, consider a recent study of the sensitivity of m_t -measurements to simulation parameters (Corcella *et al.* [2018]). Since m_t is pivotal for constraining the Higgs-mass or determining the lifetime of the SM-vacuum, it is 'of paramount importance to measure m_t with the highest possible accuracy' ([2018], p. 486).

A central problem in m_t 's determination resides in fundamental uncertainties in the energy scale of b -jets; characteristically broad jets of hadronic particles resulting from decays of heavy b -quarks, intermediate in top-decays. An alternative is to exploit decays in which the b -quark forms a B -meson, a b -flavoured hadron that then decays into other characteristic particles. In this case, jet uncertainties do not enter in the same ways, but: 'a crucial role is nevertheless played by the hadronization of b quarks into B -hadrons and the associated Monte Carlo uncertainty' (Corcella *et al.* [2018], p. 487). In order to determine the impact of changes in the

hadronization parameters on m_t , Corcella *et al.* ([2018]) performed a sensitivity analysis for relevant measurements. Surprisingly, m_t was most sensitive to the strong coupling for final state radiation in PYTHIA's (partly theory-driven) parton shower, not to any of the (essentially free) hadronization parameters (see Table 7 in Corcella *et al.* [2018]). Moreover, when inferred from certain dedicated observables, the sensitivity of m_t to simulation parameters turned out to be comparatively low overall.

Prima facie, the study delivers a solid reason to think that experimental results can be robust against changes in the simulation parameters when the observables used to infer them are carefully chosen. However, the reasoning expressed in the study has several curious features. First off, Corcella *et al.* ([2018], p. 492) avoided 'values too far from the default' and varied parameters 'up and down by at most 20% of their central values'. By itself, the restriction to a certain range of variation is nothing worrisome, and neither is, necessarily, the restriction to 20%: Standards of variation in estimating uncertainty are generally fixed by convention in HEP.

What is peculiar is the reason given by Corcella *et al.* ([2018], p. 492) for restricting their attention to 20 rather than, say, 50%; namely, that 'values too far from the default [...] might generate unforeseen changes in the Monte Carlo predictions'. The philosopher well educated in the pitfalls of reasoning should be struck by this comment. For wasn't the very reason for performing a sensitivity analysis to determine whether changes in the parameters do lead to unforeseen changes in predictions? Taken literally, this quote hence expresses a *petitio*: If the possibility of 'unforeseen changes' was used to justify confining one's attention to ranges where these are not expected, the expectation of a small impact of parameter variations would follow trivially. This cannot be right.

Now IPR could establish a rational reason for confining attention to the particular region. That is, if, given all empirical evidence, it was safe to assume that the parameters must lie in close proximity to one particular tune, varying them by only a small percentage would be justified. However, we already know that this is not true. The actual reasons for the 20%-restriction were rooted rather in the fact that the analysis was (largely) based on the Monash tune, which, to recall, was itself based on LHC data. Hence, specificity to the experimental

context grounded the restriction to a preferred region of values, not robustness across various experimental contexts.

This sort of specificity is taken to the extreme by the method for increasing accuracy ultimately suggested by Corcella *et al.* ([2018]): An *in situ* calibration of simulation parameters, that is, defining a number of ‘calibration observables’ relevant to the general measurement of m_t but rather insensitive to m_t itself, and tuning the simulation directly to m_t data. Obviously, any circularity is evaded if calibration observables are chosen such that they do not (strongly) depend on m_t .¹⁰ But: ‘it will be very unlikely for such fits to be “universal”, *i.e.*, they are not expected to be reliable outside the *very* same data sample accounted for in the parameter adjustment’ (Corcella *et al.* [2018], p. 520).

The point to be gathered is this: A high amount of PR can only be established by first calibrating parameters in a highly specific way—maybe even to one particular decay chain. This does not per se speak against PR’s relevance for establishing trust in the CSs’s result-facilitating capacity. But it forcefully underscores that it seems rather impossible to associate their parameters with a definite physical meaning, and raises the question of how else to understand them.

There is another sense in which PR can be established: In the global analysis discussed in Section 4.1, different studies also used different tunes. Hence, next to MR, this analysis also creates some amount of PR, by establishing a consistent result from several tunes. By itself, however, this ability to infer consistent results from different parameter-tunings is insufficient for securing results from error as well. An impressive example from astrophysics is discussed by Gueguen ([2019], p. 1204), where several CSs converge on a prediction for the disruption of dark matter subhaloes for a certain choice of parameters: ‘But if another region of the parameter space is scrutinized, convergence is found on another prediction, that contradicts the former’. Hence, at least one of these convergences must be spurious, produced as an artefact of the tuning.

I believe, however, that physicists in HEP are in a better position, because in the generation of QERs in HEP, such as that of the world average for m_t , several distinct tunings are used

¹⁰ For more encompassing analyses of potential circularities in HEP, see (Staley [2004a], pp. 245–47; Beauchemin [2017]; Ritson and Staley [2020]).

to generate consistent results; not just one tuning per CS, as in the astrophysics case. Taking into account also that the given tunings are not all entirely specific to the experimental context in which they are ultimately used, that is, not all generated *in situ*, this consistency between results inferred by appeal to highly specific tunes and well-validated, less specific ones gives HEP physicists a certain edge in claims to PR.

6 Trust and Its Limits

6.1 Commonality, confirmation, and two senses of trust

Overall, CR, MR, and PR together secure a rather solid basis for investing belief in experimental results, in the sense that they are probably not artefacts, and that future developments in theory, experiment, and modelling will not lead to radical changes in their central values. However, what, if anything, can and do these RAs establish about the models themselves?

Recall that a key role in RAs across models is played not only by the differences in idealization and simplification, but also by what is common to all models (Weisberg [2006]; Schupbach [2016]). The commonalities between string and cluster models, say, mostly reside in qualitative features of QCD, such as pair-creation out of the vacuum, (pre-)confinement to intermediate, short-lived hadrons ('strings' and 'clusters'), and energy–momentum conservation in the non-perturbative regime. Given that there are many shared predictions between these models, those features could give rise to 'robust theorems' (Levins [1966]; Weisberg [2006]; Weisberg and Reisman [2008]). A now-classic example is the 'Volterra property' (Weisberg [2006]; Weisberg and Reisman [2008]): that a general biocide increases the abundance of prey and decreases the abundance of predators. The common structure of ecological models predicting this is a negative coupling between prey and predators, that is, that more prey leads to more predators but more predators lead to less prey. Hence, it is a robust theorem that (*ceteris paribus*) a predator–prey system with this sort of coupling will exhibit the Volterra property.

The existence of robust theorems paves the way to what Weisberg ([2006], p. 740; emphasis omitted) calls 'low-level confirmation'; that 'certain mathematical structures can adequately represent properties of target phenomena'. This is to be distinguished from the much stronger sense of confirmation gained by inferring a result consistently from various models and em-

pirical data in concert. However, robust theorems are of the conditional form ‘If structure S is present, then, *ceteris paribus*, property P will be instantiated’, so low-level confirmation could be conceived of semi-empirically, as an induction over antecedent hypotheses: All models consistent with evidence so far (that is, that exhibited P) had certain antecedent properties (S); hence, probably any model will. In turn, this may single out the uniquely best modelling assumptions, pointing to an underlying reality.

This is a sensible account of confirmation directly from a model-based RA, compatible also with considerations by Lloyd ([2010]). A general reason for scepticism is that it is never really possible to tell how much of the hypothesis-space has been exhausted (Stanford [2006]; van Fraassen [1989]). However, there are far more specific reasons to be sceptical in this particular context, covered in Section 6.3. For now, note that what could possibly be confirmed here are very generic features of QCD, already assumed to be correct by the physics community. Hence, these robust theorems are not particularly interesting.

Non-empirical confirmation has received a new spin from the debate on theories in fundamental physics (Dardashti *et al.* [2019]; Dawid [2013]), and some authors in that debate (for example, Smeenk [2019]; Cabrera [2018]) distinguish between two senses of trust in a theory, hypothesis, or model: Trust in the sense of belief, and trust in the sense of acceptance, where in contrast to belief, ‘acceptance is context-dependent, shaped by factors other than evidence, voluntary, and exempt from demands for overall consistency across contexts’ (Frost-Arnold [2014], p. 1964).

Trusting HEP simulation models in their capacity to facilitate experimental results could mean accepting them in the voluntary, context-dependent sense elaborated by Frost-Arnold ([2014]). However, this does not imply a commitment to accepting them as candidate representations of the goings on in the beam pipe. Hence, it also does not settle the question as to ‘as what’ we may accept them.

6.2 Instrumental reliability

Observe, now, that part of the reason why PR is possible lies in the availability of various tunings. In other words: the very feature that makes these CSs suspicious as candidate representa-

tions of the goings on in the beam pipe promotes their trustworthiness in their QER-facilitating capacity. I now turn to a positive account of 'as what' we should accept these simulation models that takes this observation to heart. In this connection, note another important observation by Hasse and Lenhard ([2017], p. 102): 'Calibration [...] is used in the context of measuring instruments. Hence, [...] calibration of parameters makes models look a bit like precision instruments'. While Hasse and Lenhard ([2017]) themselves do not endorse this reading, I believe it is on the right track: The delicate status of parameters exhibited above leaves us in no better position than to regard HEP simulation models as reliable (cognitive) instruments, insofar as they establish an inferential connection between data and QERs without painting a trustworthy picture of the underlying reality.

Note that the insertion of 'cognitive' here distinguishes my position from those of Morrison ([2009]), Tal ([2012]), or Parker ([2017]), according to which the output of a CS can, under certain circumstances, be literally thought of as a measurement outcome. Understanding models as providing inferential connections, the instrumentalism I have in mind is rather closely tied to that of the traditional realism debate. However, with sophisticated instrumentalists like Stanford ([2006], p. 207), I do not commit 'the mistake of imagining that a theory [or model] is *simply* a tool for moving inferentially from some states of affairs [...] to others', but recognize that successful models and theories are also

[...] the best foundation and starting point we have not only for uncovering new and unexpected phenomena, but also for opening up new areas and paths of inquiry, and in guiding ourselves to the even more powerful conceptions of natural domains that will ultimately replace the ones we now have. (Stanford [2006], p. 207)

To see why the term 'instrument' applies at all, it is, nevertheless, instructive to consider Tal's ([2012], pp. 151–63) elaborate account of what he calls 'white box'-calibration. What is 'white boxed' here is the overall model of the experiment to which a given instrument is calibrated.¹¹ This model includes several parameters, representing the properties of the instrument itself, details about the reference data to calibrate on, the impact of background effects, *et cetera*. For an instrument indication to be informative about the value of the quantity to be measured, a

¹¹ It is important not to conflate several issues here: Tal ([2012]) cites Boumans ([2006]), who applies 'white box' to models that are sufficiently transparent to answer why-questions. But a simulation model can, of course, fail to (accurately) answer 'why'-questions about its purported target while being sufficiently transparent (*qua* instrument) to allow for the sort of calibration investigated by Tal.

‘calibration function’ must be established that delivers the instrument indications as a function of the quantity’s values, relative to the parameters of the white box-model. In a final step, a ‘backward function’ is then established, which allows to infer values from instrument indications relative to these parameters.

A subset of the parameters directly characterize the instrument. For instance, in the specification of properties of a calliper (see Tal [2012], p. 152), some, like the fineness of the scale or the length of the leg, can be considered as intrinsic to it. The remaining parameters will characterize ambient features such as properties of the reference data, experimental conditions, and so forth. Some parameters will also represent—often indirectly—a relation between instrument and ambient conditions, like the temperature of the calliper-scale, which indirectly characterizes the energy-exchange with the environment (see Tal [2012], p. 153).

Now in the calliper models, as discussed by Tal, all these properties are considered fixed. However, some instruments’ properties are flexible, and this allows adjusting the instrument to the given experimental context. For example, an ordinary pressure gauge has mechanical connections and adjustable links that can be unscrewed and repositioned, and this may be necessary depending on the ambient environmental conditions (Cable [2005], pp. 52–55). Tweaking the instrument properties in this way is formally analogous to fitting the calibration function by adjusting some of its parameters. But fitting the parameters of a simulation model may thus be viewed as adjusting an instrument’s properties to ambient conditions of use; that is, as adjusting the model to the conditions (or ones relevantly similar to these) under which it is supposed to facilitate an inference.

I believe that it is not only possible to view simulation models as instruments in this way, but even advantageous. First, if simulation models are instruments, it is perfectly understandable that each LHC collaboration comes down with its own tunes. This just means adjusting the instrument to a range of tasks, taking track of the relation between instrument and (known) ambient conditions. A procedure like this would, in contrast, make no sense if the model was to represent the goings-on within the measured system.

Second, if simulation models are instruments, then even calibrating *in situ* is nothing suspicious at all: Calibrating an instrument *in situ* means taking track of the relation between

experimental inputs and measured values in the precise context into which the instrument is embedded, and this is sometimes recognized as an advantage of *in situ* calibration on material instruments (for example, Cable [2005], p. 12). It is then also unsurprising that measurements can be associated with a smaller uncertainty, since the calibration procedure eliminates the possibility of error due to unrepresentative environmental conditions.

Third, if calibrating on the very data used for measurement means eliminating possible errors in representing the ambient environmental conditions, this also serves RA's aim of securing the result from important sources of error. It hence becomes understandable why Corcella *et al.* ([2018], p. 490) considered their procedure to not only yield 'an improvement of the accuracy' but also 'the robustness of the mass measurement'.

Assuming that these models, thus, function as instruments, in what sense can we regard them as reliable? The sense of reliability here is that when calibrated to sufficiently specific experimental conditions, they can facilitate convergent QERs in ways that do not crucially depend on any model's peculiar properties. This was fully evidenced in Section 4, and may be seen as an instance of what Parker ([2020], p. 462; emphasis omitted) calls 'reliability in a type of use'. This reliability, generating a kind of adequacy-for-purpose, is present for tool *M* when in instances of type *C*, a user *U*, following a methodology *W* under circumstances *B*, is likely to achieve purpose *P* with *M*.

As we saw, in instances of simulation-infected measurements (*C*), HEP-collaborations (*U*), following a calibration procedure (*W*) that involves relevantly similar data (*B*), can achieve measurement results (*P*) that lie in sufficient proximity to one another so as to define a(n informative) common result. Furthermore, since each measurement discriminates against different hypotheses, *H*, as to the result's potentially artefactual nature, this usage can be seen to likely serve the intended purpose. This fully establishes the sense of reliability I have in mind.

6.3 Assumption convergence: A rejoinder?

What could be countered to rehabilitate these models' representational status? As discussed in Section 6.1, a common core beyond generic QCD features could lead to 'richer' robust theorems, thus giving rise to a more meaningful kind of 'low-level confirmation'. Furthermore,

different hadronization models are not entirely independent, as mentioned in Section 3.2, and hence there is reason to believe this greater common core might exist. For instance, string and cluster models both face limitations in certain mass domains, and so lighter strings are essentially treated as clusters in string models whereas heavier clusters are essentially treated as strings in cluster models (see Dissertori *et al.* [2003], p. 171). More generally:

The accumulation of more precise data have led the models to converge to something more similar. The string model has had successively refined perturbative evolution and the cluster model has become successively more string-like. This leads one to wonder whether nature is pointing us towards a model in which the flavour mix is largely determined by the perturbative dynamics, as in the cluster model, and their distributions largely determined by non-perturbative string dynamics. (Seymour and Marx [2015], p. 309)

A possible explanation of the successful use of such converged models is hence that they at least partially capture the truth. I mentioned before that there are some general reasons for scepticism about this sort of confirmation, but in the present context, there are also quite specific reasons.

First, recall that parametrization in hadronization-modelling is applied at a level where many details remain unresolved. The success of models whose parameters reside at such a level of grain cannot be clearly attributed to their underlying assumptions:

If the model agrees with the observations against which it is tested, this could be the result of ‘compensating errors’ (namely, the model has been tuned as a whole), and if the model does not agree with the observations, one does not know which adjustable parameters (or even complete parametrisations) are wrong. (Petersen [2000], p. 269)

Peterson here refers to climate models, and one might argue that the situation is different in HEP; for it is not the success of current models but rather the failure of older models that leads to the possibility of semi-empirical confirmation.

Second, however, the *in situ* calibration for increased accuracy described above pertains to quite recent models, and each collaboration keeps producing (and using) individual tunes for specific purposes. Hence, the physical meaningfulness of converged models remains impaired by the fact that parameters will be adjusted to details that should not matter for their interpretation. I thus maintain that the actual use of HEP simulation parameters spoils the indicated possibility of semi-empirical confirmation: Being able to rule out some models across all values of parameters does not render the remaining models trustworthy beyond instrumental reliability.

7 Conclusions

I have argued that CSs in HEP can be trusted in their capacity to facilitate credible experimental results when individual results are robust against changes of simulation model and consistent results are generated using different processes to simulate, analyses making use of CSs in different ways, and even different parameter-tunings. Furthermore, by combining several results so generated, it is possible to strike a balance between security and informativeness.

However, the models involved in generating these results exhibited a curious failure of robustness regarding their parameters: Choices of tune are not necessarily consistent even across the LHC, and different tunes can in principle alter CS-based predictions non-negligibly. This delicate combination of robustness and failures thereof was used to argue that while experimental results can be trusted in the strong sense of belief, the simulation models used to generate them can be trusted only in the weak sense of acceptance as reliable instruments. The strongest reasons given for this were that hyper-specific tunes can increase experimental accuracy and that experimental collaborations each use various tunes while the physics targeted by them should be insensitive to the detailed experimental conditions.

These reservations against the models' faithful representational status could be answered by deeper modelling and parametrization at a higher grain, but whether this is actually possible remains a question for working physicist to answer. It would, of course, even then be possible to remain sceptical about semi-empirical confirmation from RA, but I believe that the remaining doubts would reside at a deeper level of philosophical debate (Stanford [2006]; van Fraassen [1989]).

Finally, the conclusions reached in this article should be of interest also for other simulation-heavy fields: Climate models, like HEP simulation models, are highly entrenched (Lenhard and Winsberg [2010]), whence independent validation is impossible; and models in astrophysics exhibit similar peculiarities regarding their free parameters (Gueguen [2019]). Furthermore, due to its overall organization and properly experimental status, HEP is subject to favourable epistemic conditions regarding the justification and use of its CSs (see Boge and Zeitnitz [forthcoming]). Hence, conclusions drawn for HEP likely transfer to fields where the conditions are arguably less favourable.

Acknowledgements

The work for this article was conducted during my employment with the research unit *The Epistemology of the Large Hadron Collider*, and funded by the German Research Foundation (DFG; grant FOR 2063). I thank the participants of the research unit's reading group for helpful comments, especially (in alphabetical order) Radin Dardashti, Paul Grünke, Rafaela Hillerbrand, Peter Mättig, Niels Martens, Gregor Schiemann, Adrian Wüthrich, and Christian Zeitniz. I have also greatly profited from comments by (in alphabetical order) Marie Guegeuen, Koray Karaca, Johannes Lenhard, Kent Staley, Michael Weisberg, Eric Winsberg, the audience of the philosophy of science colloquium at the Munich Center for Mathematical Philosophy, two anonymous referees, and an anonymous editor of the *BJPS*.

Interdisciplinary Centre for Science and Technology Studies (IZWT)

Wuppertal University

Wuppertal, Germany

fjboge@uni-wuppertal.de

References

- ATLAS, CDF, CMS and D0 Collaborations [unpublished]: 'First Combination of Tevatron and LHC Measurements of the Top-Quark Mass', available at <arxiv.org/abs/1403.4427>.
- ATLAS Collaboration [2010]: 'The ATLAS Simulation Infrastructure', *The European Physical Journal C*, **70**, pp. 823–74.
- ATLAS Collaboration [2012]: 'Measurement of the Top Quark Mass With the Template Method in the $t\bar{t} \rightarrow \text{Lepton} + \text{Jets}$ Channel Using ATLAS Data', *The European Physical Journal C*, **72**, p. 2046.
- Atlas Collaboration [2014]: 'Atlas Pythia 8 Tunes to 7 TeV Data', Technical Report *Atl-phys-pub-2014-021*, available at <cds.cern.ch/record/1966419/files/ATL-PHYS-PUB-2014-021.pdf>.

- Bähr, M., Gieseke, S., Gigg, M. A., Grellscheid, D., Hamilton, K., Latunde-Dada, O., Plätzer, S., Richardson, P., Seymour, M. H., Sherstnev, A., Sherstnev, A., Tully, J. and Webber, B. R. [2008]: 'Herwig++ Physics and Manual', *The European Physical Journal C*, **58**, pp. 639–707.
- Beauchemin, P.-H. [2017]: 'Autopsy of Measurements with the ATLAS Detector at the LHC', *Synthese*, **194**, pp. 275–312.
- Bierlich, C., Buckley, A., Butterworth, J., Christensen, C. H., Corpe, L., Grellscheid, D., Grosse-Oetringhaus, J. F., Gutsche, C., Karczmarczyk, P., Klein, J., Lönnblad, L., Pollard, C. S., Richardson, P., Schulz, H. and Siebert, F. [2020]: 'Robust Independent Validation of Experiment and Theory: Rivet Version 3', *SciPost Physics*, **8**, p. 026.
- Boge, F. J. and Zeitnitz, C. [forthcoming]: 'Polycratic Hierarchies and Networks: What Simulation-Modeling at the LHC Can Teach Us about the Epistemology of Simulation', *Synthese*, available at <doi.org/10.1007/s11229-020-02667-3>.
- Boumans, M. [2006]: 'The Difference between Answering a "Why" Question and Answering a "How Much" Question', in J. Lenhard, G. Küppers and T. Shinn (eds), *Simulation: Pragmatic Construction of Reality*, Dordrecht: Springer, pp. 107–24.
- Buckley, A., Hoeth, H., Lacker, H., Schulz, H. and von Seggern, J. E. [2010]: 'Systematic Event Generator Tuning for the LHC', *The European Physical Journal C*, **65**, pp. 331–57.
- Cable, M. [2005]: *Calibration: A Technician's Guide*, Research Triangle Park: ISA.
- Cabrera, F. [2021]: 'String Theory, Non-empirical Theory Assessment, and the Context of Pursuit', *Synthese*, **198**, pp. 3671–99.
- Campbell, D. T. and Fiske, D. W. [1959]: 'Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix', *Psychological Bulletin*, **56**, pp. 81–105.
- Collins, J. C. [2011]: *Foundations of Perturbative QCD*, Cambridge: Cambridge University Press.

- Corcella, G. [2019]: 'The Top-Quark Mass: Challenges in Definition and Determination', *Frontiers in Physics*, **7**, pp. 60–75.
- Corcella, G., Franceschini, R. and Kim, D. [2018]: 'Fragmentation Uncertainties in Hadronic Observables for Top-Quark Mass Measurements', *Nuclear Physics B*, **929**, pp. 485–526.
- D0 Collaboration [2006]: 'Measurement of the Top Quark Mass in the Lepton + Jets Final State with the Matrix Element Method', *Physical Review D*, **74**, p. 092005.
- Dardashti, R., Dawid, R. and Thébault, K. [2019]: *Why Trust a Theory? Epistemology of Fundamental Physics*, Cambridge: Cambridge University Press.
- Dawid, R. [2013]: *String Theory and the Scientific Method*, Cambridge: Cambridge University Press.
- Dissertori, G., Knowles, I. and Schmelling, M. [2003]: *Quantum Chromodynamics: High Energy Experiments and Theory*, Oxford: Clarendon Press.
- Dorigo, T. [2016]: *Anomaly! Collider Physics and the Quest for New Phenomena at Fermilab*, Singapore: World Scientific.
- Durán, J. M. and Formanek, N. [2018]: 'Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism', *Minds and Machines*, **28**, pp. 645–66.
- Fiedler, F., Grohsjean, A., Haefner, P. and Schieferdecker, P. [2010]: 'The Matrix Element Method and Its Application to Measurements of the Top Quark Mass', *Nuclear Instruments and Methods in Physics Research Section A*, **624**, pp. 203–18.
- Franklin, A. [2013]: *Shifting Standards: Experiments in Particle Physics in the Twentieth Century*, Pittsburgh, PA: University of Pittsburgh Press.
- Frost-Arnold, K. [2014]: 'The Cognitive Attitude of Rational Trust', *Synthese*, **191**, pp. 1957–74.
- GEANT4 Collaboration [2017]: *Physics Reference Manual*, Release 10.4, available at <http://geant4-userdoc.web.cern.ch/UsersGuides/PhysicsReferenceManual/BackupVersions/V10.4/fo/PhysicsReferenceManual.pdf>.

- Gueguen, M. [2020]: 'On Robustness in Cosmological Simulations', *Philosophy of Science*, **87**, pp. 1197–208.
- Hasse, H. and Lenhard, J. [2017]: 'Boon and Bane: On the Role of Adjustable Parameters in Simulation Models', in J. Lenhard and M. Carrier (eds), *Mathematics as a Tool*, Cham: Springer, pp. 93–115.
- Höche, S. [2016]: 'Introduction to Parton-Shower Event Generators', in L. Dixon and F. Petriello (eds), *Proceedings of the 2014 Theoretical Advanced Study Institute in Elementary Particle Physics*, London: World Scientific, pp. 235–96.
- Karaca, K. [2013]: 'The Strong and Weak Senses of Theory-Ladenness of Experimentation: Theory-Driven versus Exploratory Experiments in the History of High-Energy Particle Physics', *Science in Context*, **26**, pp. 93–136.
- Karaca, K. [2022]: 'Two Senses of Experimental Robustness: Result Robustness and Procedure Robustness', *British Journal for Philosophy of Science*, **73**, available at <doi.org/10.1093/bjps/axy031>.
- Kuorikoski, J., Lehtinen, A. and Marchionni, C. [2010]: 'Economic Modelling as Robustness Analysis', *British Journal for the Philosophy of Science*, **61**, pp. 541–67.
- Lehtinen, A. and Kuorikoski, J. [2007]: 'Computing the Perfect Model: Why Do Economists Shun Simulation?', *Philosophy of Science*, **74**, pp. 304–29.
- Lenhard, J. [2007]: 'Computer Simulation: The Cooperation between Experimenting and Modeling', *Philosophy of Science*, **74**, pp. 176–94.
- Lenhard, J. and Winsberg, E. [2010]: 'Holism, Entrenchment, and the Future of Climate Model Pluralism', *Studies in History and Philosophy of Modern Physics*, **41**, pp. 253–62.
- Levins, R. [1966]: 'The Strategy of Model Building in Population Biology', *American Scientist*, **54**, pp. 421–31.
- Lloyd, E. A. [2009]: 'Varieties of Support and Confirmation of Climate Models', *Proceedings of the Aristotelian Society*, **83**, pp. 213–32.

- Lloyd, E. A. [2010]: 'Confirmation and Robustness of Climate Models', *Philosophy of Science*, **77**, pp. 971–84.
- Massimi, M. and Bhimji, W. [2015]: 'Computer Simulations and Experiments: The Case of the Higgs Boson', *Studies in History and Philosophy of Modern Physics*, **51**, pp. 71–81.
- Morrison, M. [2009]: 'Models, Measurement and Computer Simulation: The Changing Face of Experimentation', *Philosophical Studies*, **143**, pp. 33–57.
- Morrison, M. [2015]: *Reconstructing Reality: Models, Mathematics, and Simulations*, Oxford: Oxford University Press.
- Muldoon, R. [2007]: 'Robust Simulations', *Philosophy of Science*, **74**, pp. 873–83.
- Orzack, S. H. and Sober, E. [1993]: 'A Critical Assessment of Levins's "The Strategy of Model Building in Population Biology" (1966)', *The Quarterly Review of Biology*, **68**, pp. 533–46.
- Parker, W. S. [2011]: 'When Climate Models Agree: The Significance of Robust Model Predictions', *Philosophy of Science*, **78**, pp. 579–600.
- Parker, W. S. [2017]: 'Computer Simulation, Measurement, and Data Assimilation', *British Journal for the Philosophy of Science*, **68**, pp. 273–304.
- Parker, W. S. [2020]: 'Model Evaluation: An Adequacy-for-Purpose View', *Philosophy of Science*, **87**, pp. 457–77.
- Petersen, A. C. [2000]: 'Philosophy of Climate Science', *Bulletin of the American Meteorological Society*, **81**, pp. 265–72.
- Ravasio, S. F., Ježo, T., Nason, P. and Oleari, C. [2018]: 'A Theoretical Study of Top-Mass Measurements at the LHC Using NLO+PS Generators of Increasing Accuracy', *The European Physical Journal C*, **78**, p. 458.
- Ritson, S. and Staley, K. [2020]: 'How Uncertainty Can Save Measurement from Circularity and Holism', *Studies in History and Philosophy of Science*, **85**, pp. 155–65.

- Schupbach, J. N. [2016]: 'Robustness Analysis as Explanatory Reasoning', *British Journal for the Philosophy of Science*, **69**, pp. 275–300.
- Schwartz, M. [2014]: *Quantum Field Theory and the Standard Model*, Cambridge: Cambridge University Press.
- Seiden, A. [2012]: 'Characteristics of the ATLAS and CMS Detectors', *Philosophical Transactions of the Royal Society A*, **370**, pp. 892–906.
- Seymour, M. H. and Marx, M. [2015]: 'Monte Carlo Event Generators', in E. Gardi, N. Glover and A. Robson (eds), *LHC Phenomenology*, Cham: Springer, pp. 287–319.
- Sjöstrand, T., Mrenna, S. and Skands, P. [2008]: 'A Brief Introduction to PYTHIA 8.1', *Computer Physics Communications*, **178**, pp. 852–67.
- Skands, P., Carrazza, S. and Rojo, J. [2014]: 'Tuning PYTHIA 8.1: The Monash 2013 Tune', *The European Physical Journal C*, **74**, p. 3024.
- Smeenk, C. [2019]: 'Gaining Access to the Early Universe', in R. Dardashti, R. Dawid and K. Thébault (eds), *Why Trust a Theory? Epistemology of Fundamental Physics*, Cambridge: Cambridge University Press, pp. 315–35.
- Staley, K. W. [2004a]: *The Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation*, Cambridge: Cambridge University Press.
- Staley, K. W. [2004b]: 'Robust Evidence and Secure Evidence Claims', *Philosophy of Science*, **71**, pp. 467–88.
- Staley, K. W. [2014]: 'Experimental Knowledge in the Face of Theoretical Error', in M. Boumans, G. Hon and A. C. Petersen (eds), *Error and Uncertainty in Scientific Practice*, pp. 39–56.
- Staley, K. W. [2020]: 'Securing the Empirical Value of Measurement Results', *British Journal for the Philosophy of Science*, **71**, pp. 87–113.

- Stanford, K. P. [2006]: *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*, Oxford: Oxford University Press.
- Stegenga, J. and Menon, T. [2017]: 'Robustness and Independent Evidence', *Philosophy of Science*, **84**, pp. 414–35.
- Suárez, M. and Cartwright, N. [2008]: 'Theories: Tools versus Models', *Studies in History and Philosophy of Modern Physics*, **39**, pp. 62–81.
- Tal, E. [2012]: 'The Epistemology of Measurement: A Model-Based Account', PhD Thesis, University of Toronto.
- Tal, E. [2016]: 'Making Time: A Study in the Epistemology of Measurement', *British Journal for the Philosophy of Science*, **67**, pp. 297–335.
- van Fraassen, B. C. [1989]: *Laws and Symmetry*, Oxford: Oxford University Press.
- Weisberg, M. [2006]: 'Robustness Analysis', *Philosophy of Science*, **73**, pp. 730–42.
- Weisberg, M. [2013]: *Simulation and Similarity*, New York: Oxford University Press.
- Weisberg, M. and Reisman, K. [2008]: 'The Robust Volterra Principle', *Philosophy of science*, **75**, pp. 106–31.
- Wimsatt, W. C. [1981]: 'Robustness, Reliability, and Overdetermination', in M. B. Brewer and B. E. Collins (eds), *Scientific Inquiry in the Social Sciences*, San Francisco, CA: Jossey-Bass, pp. 123–62.
- Winsberg, E. [1999]: 'The Hierarchy of Models in Simulation', in L. Magnani, N. Nersessian and P. Thagard (eds), *Model-Based Reasoning in Scientific Discovery*, New York: Kluwer Academic, pp. 255–69.
- Winsberg, E. [2010]: *Science in the Age of Computer Simulation*, Chicago, IL: University of Chicago Press.
- Winsberg, E. [2018]: *Philosophy and Climate Science*, Cambridge: Cambridge University Press.

Winsberg, E. [forthcoming]: 'What Does Robustness Teach Us in Climate Science: A Re-appraisal', *Synthese*, available at <doi.org/10.1007/s11229-018-01997-7>.