Synthese (2021) 199:14427–14458
https://doi.org/10.1007/s11229-021-03428-6

ORIGINAL RESEARCH

Trustworthy simulations and their epistemic hierarchy

Peter Mättig[1]

Received: 26 September 2020 / Accepted: 22 September 2021 / Published online: 2 October 2021

Abstract
We analyze the usage of computer simulation at the LHC and derive seven jointly necessary requirements for a simulation to be considered 'trustworthy', such that it can be used as proxy for experiments. We show that these requirements can also be applied to systems without direct experimental access and discuss their validity for properties that have not yet been probed. While being necessary, these requirements are not sufficient. Such trustworthy simulations will be analyzed for the relative epistemic statuses of simulation and material measurements, from which we argue that claims of their parity are unfounded. Instead, using credibility as a measure for epistemic status, and in view of the temporal and epistemic dependence of simulation on material measurements, we argue that the latter have a higher epistemic status than the former. We further argue that suggestions to qualify the epistemic status by 'defocussing' on the material connection to the target system of either human or natural experiments are misleading.

Keywords Simulations · Trustworthiness · Epistemic hierarchy · Particle physics

1 Introduction

Computer simulation (CS) has become a standard tool across almost all fields of science and triggered a lively debate about its possible epistemic implications. In this paper we address the questions of when a simulation can be called 'trustworthy' and

This paper was written with the support of the German Research Foundation (DFG) and as part of the Research Unit 'The Epistemology of the Large Hadron Collider' (FOR 2063). The author is grateful for comments by many members of the Research Group. In particular he thanks Paul Grünke for pointing him to relevant literature and for remarks in the first phase of the project, Cristin Chall for many textual comments and Michael Stöltzner for continuous encouragement. The author is indebted to Michela Massimi for valuable suggestions in particular to highlight the meaning of a trustworthy simulation. The author acknowledges several constructive comments by three anonymous referees.

✉ Peter Mättig
maettig@physik.uni-bonn.de

[1] Department of Physics, University of Bonn, Bonn, Germany

comment on the epistemic ranking of simulations and measurements. The study is based on an in-depth discussion of the experimental procedures and role of simulation at the Large Hadron Collider (LHC) (Evans and Bryant 2008). We focus on the discovery (The ATLAS Collaboration 2012; The CMS Collaboration 2012) of the Higgs boson (Englert and Brout 1964; Higgs 1964) which has also been discussed by Morrison (2015) and Massimi and Bhimji (2015). The Higgs boson was the last missing piece of the Standard Model (SM), the theory of particle physics accommodating all current experimental measurements with high precision.

The LHC at the European center of particle physics in Geneva is the world's largest accelerator complex. In the LHC, protons collide with the highest energies yet reached, and the debris of these collisions is collected in huge detectors to infer the dynamics of elementary particles. Because of its complexity and remoteness from sensory experience, the LHC is often considered an ideal case to clarify philosophical concepts concerning theorizing and experimentation. Furthermore, CS is an important part of LHC experimentation, not the least underlined by the about 20 billion events that are simulated per year per LHC experiment. Simulation at the LHC is complex, involving many layers, and is thus an excellent field to study how CSs are used in scientific practice.

We will address in more detail the relation of simulation and material measurements and derive conditions for a simulation to be trustworthy. Morrison (2015) relates trustworthiness of simulation to its possible use as a proxy for experiments: if simulations "are as trustworthy and justified as measurements, then presumably" "simulation data can replace experimental measurement in certain contexts" (p. 243), a notion we will adopt in this article. Everyday technical applications bear witness that simulations are used without dedicated prior experimental tests. In other cases material measurements are irreplaceable in the epistemic process. Thus, the question arises under which "certain contexts" simulations should be considered as trustworthy and what limits exist in applying simulations as proxies for measurements.

Virtually everyone agrees that a proper simulation has to be validated by material measurements. But to make a simulation trustworthy, 'local' validation for special set of measurement properties is not enough, but requires a 'global' validation for a large region of measurement properties, which is covered by many validations at special properties within this region. Being based on evidence, it is 'epistemic' trust that is relevant for a trustworthy simulation.

With our results we address some recent discussions on the epistemic role and ranking of simulation. While there is no doubt about the usefulness of CS and the importance of modelling to set measurement results into context, we find some claims on the epistemic role of simulations imprecise and obliterating. Although it is beyond the scope of this paper to enter into a in–depth discussions of CS's epistemic parity with material measurements, or if simulations should be considered experiments or measurements in their own right, we will comment on these claims from the perspective of the notion of trustworthiness. For example, we will critically evaluate specific arguments raised in favour of epistemic parity, and will apply widely used criteria to epistemically rank simulation and material measurements. We find no basis to claim parity, but find that material measurements are epistemically superior.

We will begin by summarizing relevant literature on validation, epistemic hierarchy, and differences between experiments and simulation. In Sect. 3, we list some general assumptions and constraints to develop our arguments. This is followed in Sects. 4 and 5 by an outline of the role of simulation in data analysis at the LHC and an in-depth example of a trustworthy simulation. After discussing several general aspects of trustworthy simulations in Sect. 6 the example will be used to derive, in Sect. 7, requirements for a simulation to be considered trustworthy. Some applications of these requirements will be discussed in Sect. 8. In Sect. 9, we will analyze specific claims of epistemic parity between material measurements and simulation, and in Sect. 10 apply general criteria of hierarchy. Finally, we summarize and conclude in Sect. 11.

2 Briefing on the philosophical debate

There is a vigorous debate on the epistemic role of simulation addressing a broad range of topics. We will restrict ourselves to those topics that have been addressed in the context of simulations of the LHC. In this section we briefly summarize points that we will take up later.

2.1 On 'trustworthy' simulations

A primary goal of this paper is to assess the conditions for a 'trustworthy' simulation. To our knowledge, no detailed discussion on this can be found in the literature.

Morrison (2015) identifies trustworthy simulations as those that can be used as proxies for material measurements. She points to the "complex methodology involved in validating simulation models" (p. 244) and attributes a separate chapter to validation. However, beyond this discussion she provides no criterion for a simulation not only being validated, but trustworthy.

Oberkampf (2019), an aerospace engineer, discusses conditions for simulations to make reliable predictions, mostly within engineering applications. He addresses the epistemic limitations of modeling and parameter choice (see also Trucano et al. (2002)) and stresses the differences between 'validation domain', inside which "several validation experiments have been conducted" (p. 89), a larger 'application domain' where one "intends to use the model from an application perspective" (ibid.), and a region outside the application domain, where one lacks knowledge and validation.

From a philosophical perspective, Winsberg (2009) associates trustworthiness with the quality of background knowledge: "background knowledge [sanctions] the trustworthiness of that model" (p. 587). He considers three kinds of background knowledge that induce trust: the soundness of the underlying theory, the physical intuition, and the technical soundness (p. 587). Trust is based on historical success. Winsberg's and Oberkampf's accounts agree if background knowledge is identified with Oberkampf's epistemic understanding and parameter choices. However, the criteria and limitations of applying these models are not spelled out, something we intend to improve on in Sects. 6 and 8.1.

2.2 Epistemic hierarchy

Trustworthy simulations are central to the debate about epistemic ranking. There is a broad agreement about the temporal priority of experiments, as expressed by Winsberg (2009): "what you need to know in a simulation ..... usually depends on things you learned from a long history of experiment and observation." (p. 591). However, different conclusions are drawn about the relative epistemic 'payoff', 'priority', 'status', 'privilege', or 'power' of trustworthy simulations and material measurements

While some philosophers argue that experiments are epistemically superior (e.g. Guala (2002); Morgan (2005); Giere (2009); Peschard (2011); Massimi and Bhimji (2015); Roush (2018)), some see them as on a par (Morrison 2015; Massimi and Bhimji 2015; Parker 2009, less explicit also Barberousse and Jebeile (2019)), and some, at least sometimes, prioritize simulations epistemically (Winsberg 2009, p. 591).

To assess these rankings, one needs a criterion to rank material measurements and simulations. However, this is only rarely spelled out explicitly. To underline the ambiguous valuation, consider Massimi and Bhimji (2015), who analyze the epistemic hierarchy of experimental measurements and simulation through the Higgs discovery at the LHC. Using two different approaches, on the one hand they 'fully endorse' the temporal and epistemic "priority of experiments over simulation" as "a claim about the reliability of our scientific knowledge and its ultimate experimental foundation" (p. 72). On the other hand, they also "side unequivocally with Morrison in defending simulations as epistemically on a par with experiments in the context of the discovery of the Higgs boson" (p. 73).

Instead of a clear criterion for the epistemic status, it is implicitly related to 'reliability' (Massimi and Bhimji 2015; Winsberg 2009), 'confidence' (Giere 2009, p. 61), 'trust' (Winsberg 2009), 'faith' (Morrison 2015). For example, Morrison (2015, p. 249) explains that her statement "CS could be considered epistemically on a par with experimental measurements" emphasizes that "the conditions under which we have faith in experimental data are sometimes also replicated in CS". Neglecting small possible, but not clearly spelled out, differences between the terminology, we will evaluate the epistemic hierarchy using the credibility assigned to simulation and experimental measurement.

There is another more implicit line of argumentation on hierarchy. For example, Morrison (2015), claims simulation's epistemic parity because "its result [is compared] to those of experiment", which entails that "the issue is one of epistemic parity of the outputs or results" (p. 226f). Thus, in essence, Morrison argues that, since experimental measurements and simulation are part of the same process, they are on a par.

Morgan (2005) and others, connect the status in the epistemic hierarchy to 'epistemic power', the ability to make "inferences back to the world" (p. 321) or the ability to obtain 'knowledge' (Roush 2018, p. 4884). Morgan argues that experiments have a stronger epistemic power than simulations since we "are more justified in claiming to learn something about the world from the experiment because the world and experiment share the same stuff" (p. 323), which seems close to the credibility criterion mentioned above. Similarly, Winsberg (2009, p. 591) connects 'epistemic power' to "[h]ow trustworthy or reliable an experiment or simulation is". He and Parker argue

that at least for their examples, solar system and black holes, respectively weather forecast simulations, have superior, respectively, at least equivalent epistemic power as experiments. Note that for their systems no experiments can be performed.

While there is no consensus on the epistemic privilege for a mature scientific field, in general, even proponents of epistemic parity acknowledge that "[e]xperiments tend to have epistemic privilege when we know very little; this is not the case in contexts in which we know enough to build reliable simulations to answer certain sorts of questions." (Parke 2014)(p. 533f).

2.3 Materiality and measurement

Proponents of the experimental priority, like Guala (2002) and Morgan (2005), point to the material and causal connection between experiment and the target phenomenon. Guala argues that "the experiment should feature the same causal processes that are at work in the real world, rather than just display some formal relation by means of a device made of different 'stuff' " (p. 69).

Morrison (2015) objects to the "focus" on material interaction because the "notion of 'being in causal contact with the system' has no relevant epistemic or ontological implications for the way the outcomes are evaluated" (p. 212). Massimi and Bhimji (2015) even deny that causal interactions are limited to experiments, and thus a "principled epistemic distinction" (p. 80) for simulation and experiment. Thus, they argue that materiality is no justification for an epistemic priority of experiments. Similarly, Parker (2009) argues for defocussing on materiality and suggests "relevant similarity is what ultimately matters when it comes to justifying particular inferences about target systems" (p. 495). Similarity, in her sense, means the sharing of formal or material properties between the object under study and the target system (p. 487).

Note that the above quotes interpret materiality as just the material interaction in experiments, i.e. with human interventions. This appears surprising, if the overarching issue of epistemic evaluation of simulation is "the problem of determining to what extent the solved simulation model represents the target system" Morrison (2015) (p. 199).

3 Terms and basic concepts of this study

We will try to clarify the concept of trustworthiness of CSs and their relation to measurements. Like Morrison and Massimi and Bhimji, we focus on the discovery of the Higgs boson h at the LHC. In view of different uses in the literature, in this section we outline and motivate our use of terminologies and concepts.

Since the main goal of the paper is to argue, if the epistemic content of a simulation is sufficient to be used as a proxy for measurements, we are less interested in the technical application of CS to numerically solve equations, needed, for example, to account for the stochastic character of physics at the LHC. Instead we will mostly address the use and epistemic importance of physics models in simulations. Furthermore, since we are interested to use CSs to learn something about the world, we do not consider

toy models. In view of debates of the epistemic relevance of simulations—also with respect to traditional models, we distinguish 'simple models' and simulations as a convolution of individual models.

We will only consider simulations and experiments that are technically correct. A simulation with a coding error is trivially inferior to an experimental measurement, and it makes no sense to epistemically rank an experimental measurement where cables are not well attached. Such error prone simulations exist, see e.g. Morrison (2015)(p. 252ff), as do loose cables, etc., in experiments.[1]

We will further compare simulations and experiments for the same problem, and do not consider that their results could be embedded in a different scientific problem. We do not deny the importance of this issue (see e.g. Guala (2002)), but it blurs the comparison of the epistemic roles of experiment and simulation. Take the example of Higgs h production by LHC proton collisions. We will compare the outcome of material proton collisions leading to a signal h with the simulation of h production in proton collisions. We will not address inferring the state of the early universe from h, which is a very different scientific problem and would require a whole string of additional material information from experiments, observations and a model.

We require an experiment to be a material intervention. However, since we consider the aim of the scientific process to understand the "causal process at work" in a target system, we include both human and (non-human) "natural" (Woodward (2003)(p. 109)) experiments in our evaluation. The latter is measured through observations. For example, in case of astrophysics, Anderl (2016) has analyzed the similarities of experimental and observational science. We agree with Woodward that experiments of both kinds are causal interventions (p. 94), one by humans, the other by natural processes, and thus both are relevant to evaluate, if 'causal contact' has relevant epistemic implications. Thus, we will replace 'experiments' by the notion of 'material measurement' that encompasses both types of measurements. In consequence, we consider it unjustified to restrict materiality to just human experiments. Furthermore we require a measurement to make a quantitative statement on the target system, in accordance with many philosophical assessments (see, e.g., Tal (2017)).

Encompassing human intervention and observation in our discussion, reflects the conditions at the LHC, and possibly a wide range of modern multi - purpose experiments. On the one hand, the LHC is a classical experiment where initial properties of interactions are set by human interventions, e.g. the total energy and kinds of colliding particles. However, the collisions of interest are primarily not those of protons, but of the quarks and gluons inside the proton. These can hardly be varied by targeted intervention to 'see how properties of interest' (Parker 2009) change. For example, it would be interesting to collide two bottom quarks at a mass of 125 GeV. However, no human intervention will be able to produce just this process. Instead, the LHC provides all kinds of parton collisions, each with a wide spectrum of energies from a few GeV to some 5 TeV. Furthermore one may be interested in the production of certain kinds of particles, say two photons or two Z's. But what the LHC delivers is a huge range of different final states. The "properties of interest" are obtained by selecting certain

[1] A loose cable was identified as the culprit leading to the claim of super - luminous neutrinos (The OPERA Collaboration 2012). For further examples of experimental claims that eventually turned out to be unfounded, see (Franklin 2013).

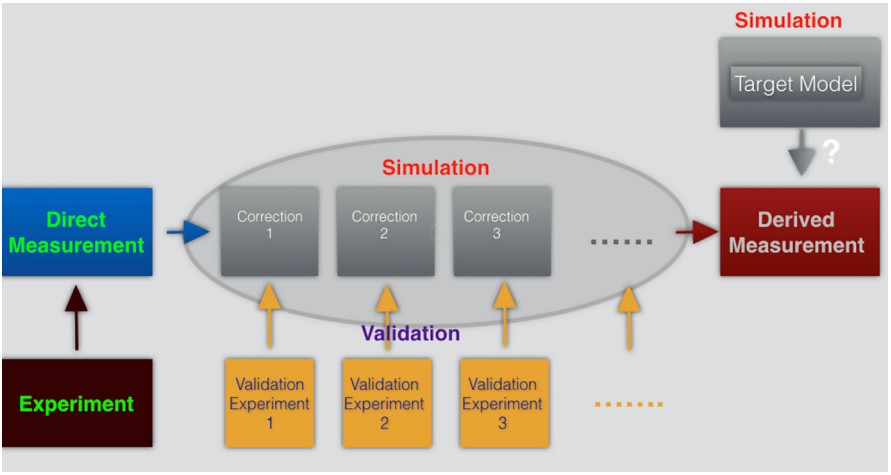

Fig. 1 Schematics of experimental measurement and connection to simulation. The direct outcome of an experiment is corrected in several steps to obtain a derived measurements of a parameter, a distribution, etc. Each of these steps is separately validated. The derived measurement is compared to the prediction of a target model

types of events, comparable to surveys of galaxies by telescopes. In consequence, the material information obtained from the LHC is a mixture of targeted intervention and observation.

4 Measurements (at the LHC)

As a basis for further discussions, Fig. 1 depicts schematically the structure of data analysis at the LHC. To characterize different stages of measurements (cp. Tal (2017), Sect.3.1) we will use the terminology 'direct' and 'derived' measurements in a fashion similar, but not identical, to Parker (2015a).

Strictly speaking, the direct measurement of h at the LHC are analogue signals in the front-end electronics inside the detector. After digitization and application of established operational procedures, they are turned into event signatures (Mättig and Stöltzner 2019), which allow physicists to classify and select events. For the Higgs discovery, the relevant signatures were events with two photons and four electrons,[2] which had certain kinematic properties. The Higgs boson was found as an enhancement around 125 GeV in the mass distributions of both two photons ($\gamma\gamma$) and two Z's (see Fig. 2):

$$h \rightarrow \gamma\gamma \tag{1a}$$

$$h \rightarrow ZZ \rightarrow \left(e^+e^-\right)\left(e^+e^-\right) \tag{1b}$$

[2] For simplicity we will reduce the discussion to electrons and not include decays into muons.

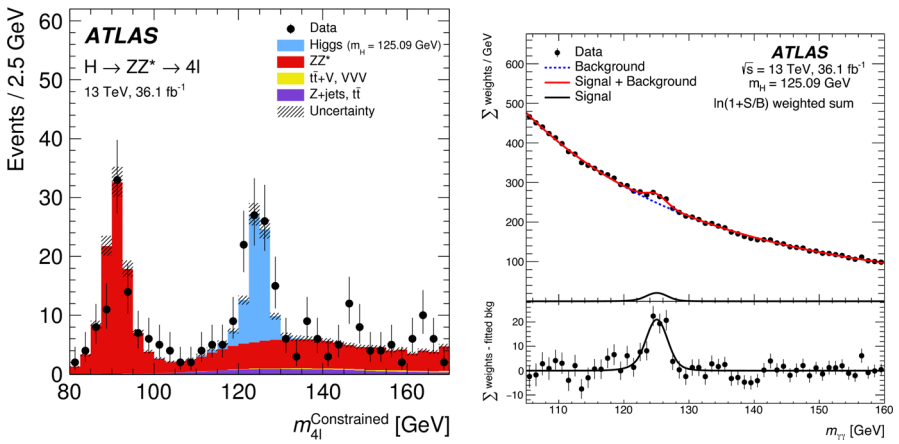

Fig. 2 Spectrum of the four lepton mass consistent with a decay ZZ, as measured by The ATLAS Collaboration (2018a), showing an enhancement at 125 GeV. The peak around 91 GeV is the Z peak. (left). Spectrum of the two - photon mass as measured by The ATLAS Collaboration (2018b) (right) with an enhancement at 125 GeV. The enhancement and the background contributions are explained in the legends

Interpreting the direct measurement of the Higgs particle h in terms of the patterns of electronic signals would be fairly clumsy. Therefore, we will instead consider the more transparent measurement of the number of signal events of Eq. 1 as a 'direct measurement'.[3] The steps between electronic signals and these signatures are part of the corrections in Fig. 1, and skipping them does not change any of our conclusions.

The derived measurement allows for precise comparisons across different experiments and to theory. It is obtained by corrections that account for detector specific distortions like efficiencies or measurement biases and effects due to the selection process, including background processes. These corrections are expressed by mathematical prescriptions that transform, step by step, the direct measurement and are combined in computer simulations (see Mättig (2019), and Boge and Zeitnitz (2020)).

The derived production rate can then be compared to a 'target model'. This could be, for example, the Standard Model (SM) prediction for the Higgs cross section, but also one of the many alternative models predicting an enhancement.[4]

Let us point to some general conclusions from Fig. 1.

- The direct measurement (blue) is essentially void of any models beyond established experimental procedures and operational definitions. In particular, it is free of any assumptions of the SM. The direct measurement is translated, via validated corrections, into a derived measurement (red), which is used to interpret the measurement with a physics model.

[3] More precisely the production rate, or, in physicists' terminology, the production cross section.

[4] Note that Fig. 1 sketches just one way to compare a direct measurement with a model prediction. Alternatively the corrections can be applied to the target model to see if the direct measurement agrees with the expected one. Also hybrid approaches are possible: the corrections are partly applied to the direct measurement and partly to the target model. For simplicity we will only use this one method. Neither procedure changes our principle discussion.

- The corrections factorize to yield the derived result. The goal is to perform a measurement that is unbiased and does not preempt the target model. Therefore all corrections are validated using complementary data from the LHC or other material sources, as well as physics models. Each correction is subject to uncertainties that bear on the precision of the simulation and that quantify the uncertainty of the derived measurement.[5] If the corrections are exhaustive, the derived measurement is thus 'valid', independent of any target model.
- The direct measurement is an *a priori* unknown part of the measurement process. Thus, while the corrections are known, the derived measurement is also unknown *a priori*. Only if the simulation and target system agree, is the target model a good representation. Claiming that the observation of the Higgs is a 'validation' of all corrections (cf. Morrison (2015)(p. 286)) is misleading and not in line with experimental practice.
- If the measurement should have any meaning for the real world, the direct material measurement is a necessary part of the schematics, and cannot be replaced by CS. Furthermore, for the derived measurement, both the direct measurement and the corrections are needed. However, their roles in the measurement process are quite different.
- Determining the derived parameter and its interpretation in terms of target models can happen at different times. For example, at the LHC derived measurements are often used months or years later as constraints on newly developed target models .
- It should be noted that the relation of derived measurement and simulation in Fig. 1 implies that the uncertainty of the derived measurement is identical with the uncertainty of the simulation. In comparing to the target model, its uncertainty may come on top of this.

5 How simulation plays into LHC experimentation

In this section, we discuss how simulations are used at the LHC and prepare conclusions about how they attain the trust of physicists. As discussed in Sect. 2.1, validation underlies a trustworthy simulation. To describe the complete validation process at the LHC is far beyond this paper, but a general overview can be found in Mättig (2019). Instead, we highlight a prototypical example of one correction, the calibration of the electron energy, which was essential for the Higgs discovery.

To arrive at the mass distributions depicted in Fig. 2, the momenta and energies of both electrons and photons[6] had to be measured. At the LHC, this is predominantly achieved with the electromagnetic calorimeter, a massive detector component where these particles are forced to interact and are eventually stopped.

To obtain the electron energy, the strength of the electronic signal in fine–grained cells of the calorimeter have to be calibrated. Initial calibration was obtained from test

[5] Along, e.g. Morrison (2015) (p. 231), we distinguish 'an accurate set of measurements gives an estimate close to the true value of the quantity being measured, and a precise measurement is one where the uncertainty in the estimated value is small'.

[6] For simplicity, and because of strong similarities in their detection we will allude to both of these as 'electrons'. Positrons are also included in this notation.

beams before LHC data taking. A higher precision is reached (The ATLAS Collaboration 2014) during data taking, where the electron energies from Z decays at the LHC are adjusted to reproduce the highly precise measurement of the Z mass from a previous experiment (Schael, S. et al. 2006). While this would be sufficient for calibration, and does not require physics models, for yet higher precision one accounts for distortions, especially due to energy losses of electrons before they enter the calorimeter. Simulation with physics models enters the stage for this final correction.

The simulation of an individual electron in the detector is described by a Markov chain of interactions in its material. Through these interactions the incoming electron induces a 'shower' of secondary electrons and photons, whose numbers are proportional to the energy of the original electron. The dominant interactions, *bremsstrahlung* and e^+e^- pair production, have been measured very precisely in the past and can be calculated with quantum electrodynamics, one of the best tested theories. Many individual interactions contribute incoherently to the showering process and, with each of the subprocesses, the number of photons and electrons increases, while their average energies are reduced. The stochastic process of electron showering is most easily solved by CS. Templates are generated with simulation for a refined adjustment of the electron energies to reproduce the Z mass.

The simulation of the electron shower is arguably the most trusted simulation of the whole chain of corrections, and can thus be used to understand the concept of trustworthiness. Apart from its use at the LHC, it has been validated in many different physical properties: for different energies, for very different experimental set-ups and materials, etc. For none of these conditions did the experimental measurement deviate 'significantly' from simulation. But even so: electron simulation at the LHC is not applied blindly, but validated by in-situ measurements. In particular, details of a shower also depend on the geometry of, and the material in and before, the calorimeter. Basically, the redundancy of the ATLAS detector allows physicists to be sensitive to potential inhomogeneities or anisotropies. For example, the distribution of energy deposits in the finely granulated calorimeter was different between simulation and material measurements, signaling poorly simulated material distributions. Further, taking advantage of the independent measurements of electron momenta in both the tracking chamber and the calorimeter of LHC experiments, angular modulations of the ratio between the two momentum measurements were observed, indicating gravitational sagging of the calorimeter. Finally, the agreement between data and simulation is constantly checked. For example, annual adjustments depend on the temperature inside the calorimeter and the conditions of data taking (The ATLAS Collaboration 2019a).

We discussed the example of electron calibration in some detail since it provides insight into how simulations are applied at the LHC. First of all, a physics model, and thus a simulation, is not mandatory in most cases. Second, the use of a physics model included in a simulation improves the precision of the calibration. Third, although the simulation may be highly trusted, it is intensively checked with (in-situ) measurements, and adjusted to account for observed deficiencies. The final authority for the correctness of simulation are material measurements. Once validated, however, the simulation is used fairly autonomously.

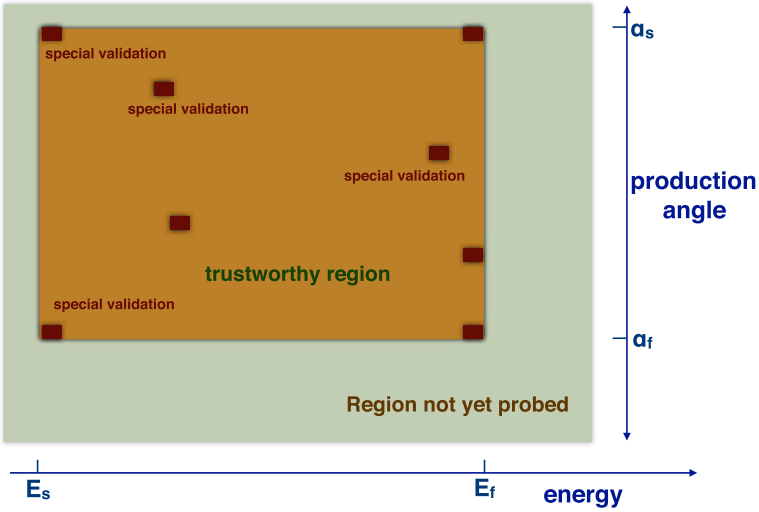

Fig. 3 Schematics of validated and trustworthy regions of a simulation. For this example, the property space is defined by the production angle and energy of measurements. The red squares indicate small regions where the simulation is validated. The brown area indicates a trustworthy region, the light green area indicates where no measurements and validation yet exists

6 From validated to trustworthy simulations

6.1 The space of measurement properties

Using the example of electron simulation, we will develop requirements for trustworthy simulations. As mentioned before, a very general and necessary condition for a simulation to connect to the world is that it is validated: subjected to "procedures for establishing whether the model fidelity is adequate for the purposes of the given application." (Murray-Smith 2019) (cf. Winsberg (2019)(p. 11)), Morrison (2015) (p. 269)). As such, validation may just apply to specific measurement properties, and not necessarily to other properties, which again would require dedicated validations. If these are needed, simulation is not trustworthy along our definition in Sect. 1 that a trustworthy simulation was defined as being applicable in a larger region of measurement properties to be usable as proxy for experiments.

We use a space of some dimension N, which we denote as the 'space of measurement properties', or short 'property space' to characterize the properties of some object under study. For example, at the LHC the space can be spanned by energy (E_i), production angle (α_i), particle types, or precision, etc.. Assume N = 2 with E_i and α_i. Similar to Oberkampf (2019) (p. 90) (see also Trucano et al. (2002)), Fig. 3 sketches such differences between merely validated simulations and trustworthy ones. In this space, the electron energy calibration is validated at many special and close–by points (red) spanning a large property space (brown) of $(E_i,\ \alpha_i) \in \{[E_s,\ E_f],\ [\alpha_s,\ \alpha_f]\}$. Here the simulation is trustworthy, a further local validation within this region is not required.

6.2 The ZZ simulation: 'Just' validated

To make the concept of a 'trustworthy' simulation more transparent, we compare the simulation of the electron shower to the simulation of another correction used for the Higgs measurement: the background to the signal at 125 GeV from continuum production of ZZ, denoted by the red area in Fig. 2 (left). To derive the number of produced Higgs bosons, the amount of continuum ZZ production has to be known. The treatment of this background (The ATLAS Collaboration 2012), is also used in Massimi and Bhimji (2015).

For the region of interest around the Higgs mass of 125 GeV, the ZZ background is simulated based on a confirmed theory with an estimated uncertainty of about 10%. The 'side bands', below and above the 125 GeV enhancement, allow physicists to adjust the outcome of the simulation by a scale factor. In the first publications on Higgs production (which were used by Massimi and Bhimji), the scale factor turned out to be close to one, and thus ZZ simulation can be considered validated in the small region around the enhancement. Since the uncertainty of the simulation was smaller than the one of the scale factor, simulation was used to estimate the ZZ rate. With more data collected in recent years, the uncertainty from the measured side bands became smaller than the one from theoretical predictions. Thus, in more recent analyses, the number ZZ events was obtained from the measurement of the side bands (The ATLAS Collaboration 2019c).

For the simulation of the ZZ continuum to be trustworthy, it has to be usable as a proxy for a large region. High mass ZZ systems, however, are increasingly accompanied by quarks or gluons. While their production can be described by the well confirmed theory of strong interactions (QCD), technical difficulties allow only an approximation (see, e.g., The ATLAS Collaboration (2019b)). For example HERWIG (Bähr, M. and others 2008) or SHERPA (Hoeche et al. 2013) models invoke QCD inspired concepts, requiring several parameters that are tuned such that the models describe the measurements. The distributions at high ZZ masses or high transverse momentum predicted with these models differs by some 20–40% from the measurements—more than the presumed uncertainties of each model. Thus, the ZZ simulation, even if validated at low masses, cannot be applied for high masses without further validation, it cannot be used as a proxy for a measurements.

6.3 What trustworthy simulation should achieve

The trusted simulation of electron showers on the one hand, and the ZZ simulation that requires validations dedicated to a specific physics problem, will be used to develop conditions for 'trustworthiness'. Before discussing these, let us clarify how we will use validation and what a trustworthy simulation should achieve.

For the following, we refrain from detailing validation procedures and simply assume validations have been performed. Procedures have been addressed in a large amount of literature—see for example the extensive survey in Beisbart and Saam (2019). However, we specifically emphasize the importance that scientists place on validation methods that are as close as possible to material measurements. The goal is

to remain independent from involved models – see the example of the electron energy calibration. This goal is in general underexposed in the literature on validation but will be important for our later discussions. Our analysis is thus not affected by potential epistemic concerns around methods of validation like the meaning of tuned model parameters (e.g. Lenhard and Hasse (2017)) or robustness (e.g. Gueguen (2020)).

Validation has also to be seen in the context of what trustworthy simulations are needed for. We see at least two purposes: for one a, frequently technical, application to predict a state and, for two, to test a hypothesis, the target model of Fig. 1. Examples for these two classes are how to decelerate a Mars–rover before touch-down or the weather in a city at noon time, respectively, to test, if the observed enhancement agrees with the SM Higgs hypothesis. For these different purposes of trustworthy simulations, target models attain different meanings, as do requirements for validation. For applications, target phenomena, like the environment of the landing site, or the topology of the city for which the weather is predicted, also have to be validated. In case of hypothesis testing, the target model is per definition not validated.

Both hypothesis testing and applications tightly connect simulation to the material world. In consequence, all our conditions for a trustworthy simulation match up with material measurements and a trustworthy simulation is a reliable one. Following Winsberg (2010) (p. 120ff), a trustworthy simulation need not commit to truth or reality. If or if not one assigns truth to either the corrections or the global outcome of the simulations, does not affect the conclusions of this paper. However, let us emphasize that for a realist to claim the reality of the target, agreement between the results of a trustworthy simulation and the target model is necessary—even it is probably not sufficient to reject an anti–realist stance.

7 Conditions for trustworthy simulations

In the following we derive conditions when a simulation is trustworthy. As will be outlined, these conditions are not fully independent from each other.

7.1 A quantifiable simulation

While qualitative simulations may provide important information and support in-principle arguments, they are only heuristic and cannot serve as a proxy for a measurement. Instead, since a measurement is a quantitative statement and validation is a quantitative characterization of the agreement between model and measurement:

Necessary Condition I
A trustworthy simulation must have a quantitative result.

This condition is all the more important, if several convolved models are used, as is typical for simulation. In this case for each of the different models a definite initial state is required.

7.2 The trustworthy property space

The very notion of trustworthiness requires a simulation to be trusted in a large property space. Ideally, one would have an infinite number of different points i in the property space, where simulation and experiment have been shown to agree. Formally, trustworthiness is thus

$$\lim_{i\to\infty} (\text{validated condition})_i \;=\; \text{trustworthy simulation} \tag{2}$$

In actual practice i will always be finite and will not cover the whole property space. We therefore relax the requirement to

Necessary Condition II
A trustworthy simulation must significantly agree with many material measurements in a large region of property space.

At the bottom of why scientists suffice with a finite number of measurements is the assumption that validated simulations at special points can be smoothly interpolated. The justification for such an interpolation grows with i and the closeness of validated conditions. In general, the number of material measurements will increase with time and they are extended into new regions of the property space, which can become trustworthy. Trustworthiness is thus a function of time and develops with more material measurements and better techniques (see, e.g., Parke (2014) (p. 533)).

We are aware that Condition II is somewhat vague, by not specifying, for example, 'how many' measurements are needed or how 'significant simulations must agree' and such vagueness will also apply to the following conditions. While we will provide some guide lines in the following discussion to mitigate the vagueness, we will discuss in Sect. 8.4 that general and precise quantitative statements are not meaningful.

7.3 The background knowledge

Winsberg (2009)(p. 589) emphasizes background knowledge, "reliable principles of model building", to trust simulation (see Sect. 2.1). In a simulation with convolved simple models, such background knowledge has to apply for each individual model or correction to interpolate between globally validated observables. Thus, trust in a simulation requires more than just global validation. In light of our previous discussion this background knowledge can be a well established theory or the direct application of complementary data or at least of data–driven arguments.[7]

Necessary Condition III
For a trustworthy simulation, each of its convolved models must be trustworthy. Each one must agree with material measurements, ideally covering the whole trustworthy region of the property space. Furthermore all models together must cover all important corrections of the material process (be exhaustive).

[7] At the LHC the term data–driven denotes a validation procedure that is largely determined by data, although not completely free of elements of the simulation. It is thus different from the terminology of data–drivenness denoting Big Data analyses.

Conditions II and III are related, but not redundant: if each convolved models would be validated for identical properties, Condition II would follow from III. However, this is rarely the case and would be too restrictive a condition. On the other hand, a validation of each simple model for some properties inside the trustworthy region is required to assign trust in interpolations by reducing the potential of compensating effects (cp. Sect. 8.1).

In a strict sense, the requirement of exhaustiveness is impossible for a simulation to meet, since simulation always invokes abstraction and idealization. Scientists account for this by assigning a systematic uncertainty (see below). Thus exhaustiveness means that at least those corrections have to be included, whose omission would lead the simulation to disagree with validating measurements by more than the simulation's uncertainty. We will return to this problem in Sect. 8.1.

7.4 Technical means

For the example of ZZ production, physicists apply highly confirmed theories of electroweak and strong interactions. However, even though the fundamental Lagrangian is well known, the observable distributions can only be calculated using approximations. This is different from the electron shower, which relies on electromagnetic interactions which can be rigorously solved with only a small uncertainty.

Thus, a well–established background theory is not enough to render a simulation trustworthy.

Necessary Condition IV
For a trustworthy simulation the computational means have to exist to solve the underlying equations precisely.

Appropriate parametrizations and approximations are in particular important if physics model cannot be derived from the first principles of a theory. For example, if theory leads to singularities even a numerical simulation would fail for an infinitely fine grid. Often a finite result can be obtained by integrating over the singularity, thus recurring to a coarser grid to find an appropriate parametrization. Winsberg (2009)(p. 587) refers to those as"computational tricks" or the "the simulationists' intuition" (see also Winsberg (2010) (p. 127ff)).

7.5 A reliable estimate of uncertainties

In practice, omissions of details and the effect of idealizations are taken care of by assigning systematic uncertainties (Cranmer 2015; Staley 2018). Given the unavoidable disagreement between the idealizing simulation and measurements, without such uncertainties there is no measure if a simulation could count as a proxy.

Necessary Condition V
A trustworthy simulation has to include a well justified estimation of its systematic uncertainties.

Such an estimation is best obtained using standard validation methods. Condition IV is necessary for Condition V, but not sufficient: while a systematic uncertainty due to a specific model is justified from the solutions of its underlying equations, additional models with different ways to approximate may exist. Thus an uncertainty is sometimes assigned by comparing different solvable, but approximate, models of the same process.

One should note that there is no requirement on how large the systematic uncertainty should be (see Sect. 8.4).

The following two conditions define limits of a trustworthy simulation. They can be derived from the previous conditions, however, because of their importance and since they have not been strongly discussed before, we list them separately.

7.6 'Trustworthiness' within precision

Uncertainties of simulation restrict the property space of trustworthiness. Take a physics model used in a simulation with a parameter of value of y and uncertainty δ, while the parameter can be measured as x with an uncertainty of ϵ. Further assume y and x agree within uncertainties. If

$$\delta \lesssim \epsilon \tag{3}$$

the simulation result could be meaningfully applied and might serve as a proxy for experimental measurements. However, if

$$\delta \gg \epsilon \tag{4}$$

simulation would impede the best derivation of the target parameter. In the region.[8]

$$[x - \delta,\ x + \delta] \tag{5}$$

simulation is blind, cannot contribute any trustworthy estimate and thus cannot replace a material measurement. This is the essence of our discussion of the ZZ background under the Higgs peak in Sect. 6.2. The small sample of early data corresponded to the condition of Eq. 3. With more data, however, it came to reflect the condition of Eq. 4.

This leads to a bound for trustworthy simulations.

Necessary Condition VI
Within its uncertainty range, the simulation cannot provide any trustworthy statement. A trustworthy simulation can only be meaningfully applied in the property region outside of the simulation's uncertainty.

Condition VI is also a statement on the simulation's potential to confirm or reject a hypothesis. If a target model A predicts a value that is different from the measurement by several times the simulation's uncertainty, model A can be disconfirmed.[9] However,

[8] For simplicity we neglect the probability distribution of the estimator.

[9] More precisely, the uncertainty is given by the convolution of the simulation's uncertainty with the one of the target model.

if two target models B and C predict values that agree with the derived measurement within the simulation's uncertainty, no statement can be made as to the relative validity of C and B.[10]

7.7 Probing new property regions

As a further clarification, we add a seventh condition that follows from Condition II.

Necessary Condition VII
A simulation can only be trustworthy in property regions that have been probed by material measurements. Outside of the probed region, a simulation cannot be considered trustworthy.

Simulation can only provide hypotheses about untested material properties. Extrapolations of a simulation into a new property region are thus *per se* uncertain and require special validations. In Fig. 3 this corresponds to the green region. We will discuss this in more detail in Sects. 8.2 and 8.3.

8 Some considerations on trustworthy simulations

We have derived seven conditions that we consider jointly necessary for a simulation to be 'trustworthy'. If one of these is not fulfilled, the simulation cannot be used as proxy for experiments.

8.1 Necessary but not sufficient

While we argued that these requirements are necessary, they are not sufficient. Simulations may agree with the material measurements at validated points and fulfill the requirements of trustworthiness, but the unavoidable abstractions and idealizations of models, a false interpolation, or an overlooked correction may lead to a wrong expectation. With denser and more precise validations, such complications will become less likely, but cannot be definitely excluded. Let us clarify this with two examples.

The large but finite amount of special validations (see Eq. 2) requires interpolation between the validated points in property space. The model may assume smoothness for interpolations, which, however, is not necessarily fulfilled. For example, mechanical, electrical, or quantum-mechanical resonances appear in special conditions. If they are not properly included in the simulation, they lead to deviations from a smooth distribution and the naive interpolation.

Furthermore a physical process may be an "unknown unknown" (Rumsfeld 2002) that is not represented in simulations. Large systematic uncertainties can hide its effect,

[10] To give an example: if an alternative model of mass generation, say a composite Higgs, predicts a rate of Higgs production that is higher than the measured one by several times the uncertainty, the simulation is trustworthy enough to disconfirm the composite Higgs model. If models like the SM Higgs or a supersymmetric Higgs predict rates that differ only within the simulation's uncertainty, the simulation is not trustworthy enough to discriminate between these two.

but the door to misrepresentations is open. Take an example where processes cancel each other: an observed distribution $f(x)$ should be described by a model A such that within uncertainties $f(x) \sim \alpha \cdot a(x)$. However, there is a further unknown and hidden effect B, and the correct distribution would be $f(x) \sim \alpha' \cdot a(x) + \beta \cdot b(x)$. While the model $\alpha \cdot a(x)$ would be validated, it is wrong and, for example, if $a(x) \sim b(x)$, the accurate model parameter would be $\alpha' = \alpha - \beta$. The contribution of $\beta \cdot b(x)$ may be hidden by assigning a large uncertainty to $\alpha \cdot a(x)$, but this may render the derived result not only unnecessarily imprecise but inter- and extrapolating the simulation may be flawed, once a sensitivity is reached to distinguish the x dependences of $a(x)$ and $b(x)$. Such ambiguities are usually resolved by higher precision, extending the property region, or comparing the desired parameter α with different methods and different auxiliary hypotheses.[11] However, this is not always possible in the short run.

8.2 Trustworthiness for new observations

Let us discuss condition VII in more detail. Before addressing new observations outside the trustworthy region, we comment briefly on new observations within.

More data allows a higher statistical sensitivity, the observation of a new effect therefore takes some time—as in the case of the enhancement at 125 GeV, which was discovered only three years after turn-on of the LHC. Here, all corrections in Fig. 1, needed to translate the direct to the derived measurements, were within the trusted property space. Although, as Morrison (2015)(p. 245) pointed out, simulation could not have claimed the existence of the enhancement, once the LHC experiments had observed it, they could use the trustworthy simulations to derive the number of signal events. It is an instance of the discussion of Sect. 4: the measurement of the enhancement at 125 GeV is valid, even though at first its underlying physics was uncertain (Chall et al. 2019). In a next step it could be compared to different target models.

Trustworthiness is more of an issue, if measurements are performed outside of the as yet confirmed property space. Frequently, scientists just extrapolate a trustworthy simulation, but in virtue of not meeting conditions II and III, the simulation in the extended space is not trustworthy *per se*. While this is a general problem, it becomes particularly virulent where the expectation deviates from the measurement. Such a deviation may signal an inappropriate target model with important epistemic consequences, but it may also be due to simple and fairly uninteresting inappropriate corrections. Dedicated validations in these new regions are required, though they may not be possible in the short term.

Instead, typically several justificatory arguments for corrections are invoked. For example, rather than validating each individual model inside a simulation, a combination of several corrections can be validated, agreements of details between simulation and material measurements can be performed, and the consistency with other measurements can be studied. To return to the calibration of the electron energy at very high energies, although no reference process like the Z decay may exist, the simulation

[11] Thus we do not argue for a Duhem - Quine - like underdetermination. Scientific practice provides means to resolve such cases.

can be justified by comparing the lateral and longitudinal energy deposition in simulation and material measurements, or measuring quantum electrodynamics processes at high energies. Other methods for multivariate analyses are, for example, based on testing complementary distributions (see Mättig 2019, Sect. 26.10.2). In actual practice, the different corrections used in the simulation are weighted according to their assumed importance and quality of extrapolation. For example, if a correction is of minor importance in the trustworthy region, it is assumed to also have a small effect in the extended property space. Also, extrapolating predictions for the electron shower appears well motivated given the principles of its simulation, while corrections like gluon or quark radiation may be more difficult to evaluate. Simply extrapolating simulations is not necessarily wrong. However, it leaves a shadow of doubt, is tentative, and requires special justifications, if a full validation is not possible.

Let us discuss two recent examples in particle physics, where measurements probed new regimes and found unexpected results. The first one is an indication of an enhancement in the di-photon mass distribution at 750 GeV (The ATLAS Collaboration 2016; The CMS Collaboration 2016). Although not statistically significant, the indication aroused a lot of theoretical and public interest, before it disappeared with more data. Such an enhancement is difficult to explain by improper corrections, and, for example, would appear even if the calibration of the electron energy were wrong. Still, several detailed, special studies have been performed of possible photon misidentification and photon isolation from other particles.

The other example is a measurement of an apparent deviation at the tails of a smooth distribution. In 1995, a higher than expected yield at the highest jet energies was found. The deviation grew continuously with energy, suggesting to some that quarks could be composite (Dorigo 2016). However, since such a deviation could easily be due to inappropriate corrections, the derived measurement raised significant doubts. By cross checking with other measurements, the apparent deviation was identified as due to improper modelling of the quark content in protons.

These two examples underline that extrapolations into new regions are in general less trusted than interpolations in established property spaces. In the latter case there are densely individually validated properties, which, by definition, do not exist in the new regions. However, as discussed, both have epistemic risks. Furthermore the level of trust depends on the kind of measurement. Means exist to tentatively justify the extrapolations by material measurements, among them importantly consistency with other independent measurements, even though a direct validation may take some time.

8.3 Trustworthiness for remote material systems

The previous discussion connects to simulations for remote material systems (sometimes called 'inaccessible target systems'), on which no experiment can be performed. Since experiments are not possible, these systems are frequently considered witness for a superior role of simulations. Examples include the solar system Winsberg (2009) and galaxy formation Morrison (2015). To arrive at observational statements, the structure of data analysis is similar to the one depicted in Fig. 1 for the LHC measurement. The

validating measurement in case of a remote system are multiple targeted observations. Do trustworthy simulations exist for remote systems?

Winsberg (2009)(p. 591) answers with 'yes', at least for the solar system: "because in such a case the relevant background knowledge, our ability to build good, reliable models is virtually unassailable". Indeed, quantitative predictions of the orbits of planets are possible, and a huge number of material measurements agree with these predictions. The background theory of Newtonian gravitation is highly confirmed, and technical means exist to solve the Newtonian equations, so systematic uncertainties can be reliably estimated and boundaries for the trustworthiness of simulation can be derived. Thus, we find all requirements for a simulation to be trustworthy are fulfilled.

To be clear about the kind of material information needed, let us remind ourselves why we are confident about Newtonian gravity, at least at the scale of the solar system. First there are earth - bound experiments like Galileo's tilted plane or the Cavendish experiments. While these experiments test gravitation at small distances, it is trusted at much larger distances beyond the reach of experiments in Newton's time. Newton did infer the $1/r^2$ dependence of the gravitational force from observations of planetary orbits, rather than experiments. Last, but not least, our confidence is based on the ability of Newtonian gravity to predict new planets to account for observed deviations.

As another example, consider simulation for galaxy formation, which Morrison (2015) considers trustworthy and elevates to "experimental knowledge", "simply because" experiments are impossible (p. 214). Morrison interprets experimental knowledge as legitimation "to claim that simulations can measure theoretical quantities" (p. 199). Our understanding of galaxy formation is characterized in the words of astrophysicists Somerville and Davé (2015) by (a) enthusiasm about the amount of material measurements: "[w]e are truly living in a golden age of facilities and databases for studying how galaxies formed and evolved."(p. 52); and (b) by qualitatively good, but quantitatively insufficient simulations: "overall we would give today's suite of galaxy formation models a passing grade" (p. 101). Thus, while simulating galaxy formation has made enormous progress over the past decades and provides much information and insight – including, importantly, limits of the current understanding, simulation of galaxy formation cannot serve as proxy for material measurements. On account of not meeting several conditions (at least II and III) it is not trustworthy. Neither can it measure a theoretical quantity.

Human interventions are impossible on the remote systems. But this does by no means imply that they are void of material measurements. For one, simulations of remote systems are largely based on results from earth - bound experiments that are extrapolated to new properties. For example, simulation of galaxy formation has to include general relativity, hydrodynamics, nuclear processes, etc., so most of its ingredients are thus applications or extrapolations of validated earth - bound experiments. Secondly, there is a large range of observations. Models of galaxy formation build upon distribution functions of galaxy properties like luminosity and wavelengths over a large frequency range, yields of galaxy types, etc.. In this respect the trustworthiness of simulations of remote systems are based both on human and natural experiments. Although human experiments are more easily controlled, the vast variety and quality of material observations would make a distinction of validation for the two kinds of material measurements artificial.

The simulation of the solar system may be taken as an instance of simulation for application. On the other hand, simulation of galaxy formation is geared towards hypothesis testing in the sense of probing if all elements of this complicated system are correctly accounted for, but also since it introduces hypotheses, which are not well understood, particularly Dark Matter. Indeed, Dark Matter is an example, where (as of today) no earth–bound information is relevant enough to allow for a trustworthy simulation.[12] Instead the only source of measurements on Dark Matter are astrophysical observations. Thus Dark Matter models appear as target models in simulations, and inferences on Dark Matter depend on the trustworthiness of all other corrections in the simulation together with their corresponding uncertainties. This complicates trust in such context.

8.4 Trustworthiness depends on epistemic goal

It is tempting to identify the systematic uncertainty of a simulation as a measure of trustworthiness. However, the uncertainty defines only limits of the trustworthy property range. Even a simulation with a large uncertainty can be trustworthy, though its range is then strongly restricted. Whether such a simulation is considered trustworthy 'enough' depends on the epistemic goal. Such a dependence is also one reason for the vagueness of some of the requirements.

Take Winsberg's solar system example. Restricting the background knowledge to Newton may be sufficient for his purposes, but it is famously insufficient to describe the Mercury perihelion, which is remedied by General Relativity. Furthermore, given the material information and the gravitational laws, there is no guarantee that the solar system remains stable for a very long time. Winsberg's ability to infer to the world, therefore, depends on his desired precision and time scale. In any case, the example shows that the trustworthiness of a simulation always applies to a limited property space.[13] This may be sufficient for certain goals, but may fail if tighter requirements are needed. It is not meaningful to quantify requirements for all purposes.

Similarly, for Condition II we did not define 'how many' material measurements are required or how dense these should be. Indeed, it does not appear meaningful to define a general criterion, since it depends on the quality of interpolations. After many attempts, we strongly assume that the law of fall $s = 1/2 \cdot gt^2$ (s being the height of fall, t the time and g the gravitational constant, with friction and other disturbing factors neglected) can be smoothly interpolated. It appears that just a few measurements would suffice. This is not the case for systems with highly non–linear effects and resonances, even if all important laws are known. The London Millennium Bridge was closed after observing unforeseen resonances. A smooth interpolation due to stability measures was not reliable—more validating measurements would have been needed.

Such vagueness applies to (almost) all our conditions, and, in fact, scientists themselves sometimes have different opinions on when exactly the conditions for

[12] Our ignorance becomes apparent by the huge allowed possible mass range of 90 orders of magnitude.

[13] Roush (2018) argues for epistemic priority of experiment but stating that 'in the extreme case, if there are no ... unknowns' (p. 4896), the superiority claim would not hold. The argument is not wrong but unrealistic. Our accounting for systematic uncertainties generalizes Roush's argument and makes it more precise.

a trustworthy simulation are fulfilled. This does not invalidate the generality of the criteria.

9 The claim of epistemic parity

Trustworthy simulation are sometimes claimed to be on a par with material measurements. We start analyzing specific claims in the literature, before evaluating the hierarchical status using general criteria in the next section.

9.1 The role of simulation in measurements

Morrison (2015) (p. 287ff) devotes a whole chapter on the role of material measurements and simulations at the LHC. Addressing all her claims is beyond this paper, but here is a brief summary of her line of argument. Her assessment is that "the mass measurement associated with the [Higgs] discovery is logically and causally dependent on simulation" (p. 288). Doubting "the very distinction between experiment and simulation; the latter is simply an integral part of the former" (Morrison 2015)(p. 316), she considers simulations as measurements, even to the extent that they can confirm an experiment. Thus, simulations are epistemically on par with material measurements and materiality does not play a justificatory role for acquiring knowledge.

Before arguing why we consider Morrison's claims mistaken, let us just add that we do agree with her that, by comparison with "actual experiments", the result of simulations are "a way of determining what exactly the experiment has shown" (Morrison 2015)(p. 245). Almost by definition, understanding, for example, the enhancement at 125 GeV as a representation of the mechanism to generate elementary masses requires a model. But this is only part of how Morrison sees simulation.

We depicted schematically in Fig. 1 that several corrections are needed to proceed from the direct measurement to the derived one. For the Higgs measurement we exemplified two of them, electron calibration (Sect. 5) and the background beneath the enhancement at 125 GeV (Sect. 6.2), motivating that a physics model[14] is not needed for either correction. In contrast to Morrison's assertion, virtually every single correction to determine the number of events in the enhancement at 125 GeV can be derived from data alone, using templates obtained directly from material data.[15] Thus, the mass measurement of the enhancement cannot depend logically and causally on simulation, as already suggested by Massimi and Bhimji (2015)(p. 81). While physicists strive to reach independence of physics models, they compromise in actual practice to use models in the simulation because of their convenience for inter- and extrapolations, and thus their adaptability to special conditions (cp. Sect. 5).

Morrison's claim that "direct comparison between object and target system based on materiality do[es] not play a justificatory role." (p. 237) is also in conflict with

[14] Some may consider a dimensions like GeV or the assumption that the same Z particle is produced in e^+e^- collisions and $pp \rightarrow e^+e^- + X$ a model. However, the models that Morrison (2015) (p. 287ff) discussed appear to have a very different quality.

[15] Such corrections with data alone would still be performed with Monte - Carlo simulation because the stochastic nature of LHC processes require numerical methods.

scientific practice. Each of the above corrections receives its justification only through direct comparison with material target systems, either by using data directly, directly adjusting models to data, or, if models are trusted, they are so by virtue of agreeing with a material target. These procedures may be concealed by the apparent complexity of data analysis, but complexity does not affect an epistemic quality. Morrison tries to circumvent the justification by materiality by demanding a 'direct' comparison, and interpreting physics models as a "representation of materiality" (p. 223). But such a representational function does not live a 'life of its own' but can represent only insofar as a model is grounded in material measurements. Thus, models, being a function of materiality, can hardly deny the justificatory role of materiality. This is underlined by our discussion in Sect. 5, where we showed that even a trustworthy simulation like the electron shower is scrutinized and modified to accord with material measurements.

Denying materiality a justifying role, Morrison continues to elevate simulations to "measurements" themselves, such that their "outputs [are] epistemically on a par with experimental measurements" (p. 205). The Higgs measurement already disagrees with this statement, but let us consider Morrison's example of the measurement of quark masses (p. 237). (Light) quark masses are derived by starting from material measurement of hadron masses, which are used as input for involved QCD calculations on a lattice that are most easily solved numerically by Monte Carlo simulations. For example, the hadron η_c mass was measured as 2.985(3) GeV (Amsler et al. 2008), from which a QCD - lattice calculation derived the mass of the charm quark as 1.273(6) GeV (McNeile, C. and others 2010). To address if the QCD calculation is a measurement, let us compare it to the traditional material measurements. What happens in a material measurement is that a detector records a definite signal that is then analyzed using operational procedures and models (cp. Fig. 1) to obtain an unambiguous derived result. The QCD model does not record anything but is a collection of equations which may be translated into a computer algorithm. It becomes operational once some scientist feeds it a starting value. But this value can be completely arbitrary, and the model responds with a value for the quark mass with no relation to the world. To become a measurement of the world, the input to the model has to come from a material measurement. But then simulation can be at most *part* of a (derived) measurement. This, however, means simulation is not in itself a measurement—as an engine is part of a car, but not the car itself. Thus Morrison's interpretation of models as a measurement cannot be upheld.

Interpreting simulations as being on a par with material measurements, Morrison (2015)(p. 245) takes a further step. She claims "simulations frequently function as the confirmation of experiment" by "deduc[ing] the nature of the debris produced in [particle] collisions", which are "crucial features of the discovery process". Now imagine that different simulations predict different experimental outcomes, say a composite Higgs (Model A), some variant of supersymmetry (Model B) and the Standard Model (Model C) all predict an enhancement at 125 GeV, but with different properties. The LHC data eventually lead to a high confirmation level of Model C which, according to Morrison, would confirm the experiment. But, at the same time, on her account, Models A and B would disconfirm the experiment. It would be up to the model preference of a physicist to confirm or disconfirm an experiment. This appears inconsistent and is certainly at odds with scientific practice.

Since simulations fail to be measurements, Morrison's claim of parity with material measurements cannot be maintained.

9.2 Does simulation causally interact?

Proponents of epistemic priority of experiments motivate their stance by the material connection between cause and target in experiments that is lacking in simulations. Massimi and Bhimji (2015) denote this argument as 'causal interaction claim' (CIC) and argue against the special role of experiments. They identify three types of CIC. Pars pro toto we will discuss CIC_1: "[e]xperiments involve direct causal interactions with the target system when a physical quantity is calibrated by direct comparison with observed data." (p. 74). An example is "comparing experimental data about the hydrogen spectrum with the known Balmer series" (p. 80). Arguing along the estimate of the ZZ background under the Higgs boson (see Sect. 6.2), they claim that scaling the simulation result to the observation is performed "in an analogous way". Thus they "question the conclusion that it is in fact this material feature that bears the burden of the job when it comes to measuring and experimenting"(p. 75).

Let us analyze this rejection using the scaling of the electron energy discussed in Sect. 5 instead of the atomic Balmer series. The electron energy of target experiment (LHC) is scaled to agree with the energy of the calibrating experiment (LEP), schematically

$$\text{Calibrating Experiment} \;\rightarrow\; \text{Target Experiment} \tag{6}$$

To use the Z mass, one does not have to understand the Standard Model, one simply has to know that the Z mass is 91.14 GeV.

In contrast, while formally the ZZ simulation is also scaled, to determine the size of ZZ background requires not only knowledge of the Z mass, but the whole machinery of the Standard Model, including very involved calculations. Then, however, physicists have to validate the calculation using material measurements to derive a scale factor for the calculation (see Sect. 6.2) Only then can the result be applied to the target. Schematically

$$\text{Simulation} \rightarrow \text{Calibrating Experiment} \;\rightarrow\; \text{Simulation Adjustment} \;[\rightarrow\; \text{Target Experiment}] \tag{7}$$

Thus Massimi and Bhimji's calibration of simulation works through material measurements and only underlines that "material features bear the burden", exactly what they want to reject. Since CIC_1 cannot be rejected, the claim of undercutting the epistemic priority of the experiment is not justified.

9.3 Material information and epistemic 'power' of simulation

Morgan argues for higher epistemic power for experiments than simulations using the criterion of justification level (see Sect. 2.2). Parker (2009) and Winsberg (2009) instead assign an equal, or even superior, epistemic power to simulations. Since they

are not explicit about their measure of epistemic power, we will just address some of Parker's and Winsberg's arguments.

Parker argues for a reduced epistemic role of laboratory experiments by replacing the justificatory role of materiality for inferences with similarity (p. 495, cp. Sect. 2.3). However, to infer with similarity, in the first place one has to know if, respectively, how similar a model is to the target. Parker (2015b) discusses the problems of providing a generally applicable measure, but much of her qualification, e.g. her requirement that a 'similar' model is able to address new questions (p. 275), points to the concept of a trustworthy model. As pointed out in Sect. 6 the conditions for a simulation to be trustworthy are ultimately founded on material measurements—how then can 'similarity' take the role of materiality as what "ultimately matters" for inference?

Parker details her objection against materiality assessing the 'same stuff' argument of Guala (2002) (p. 69) (see Sect.2.3). She argues "even when experimental and target systems are made of the same materials, there may be every reason to think that the systems are not similar in all of the relevant respects" (p. 494). Indeed it is not straightforward to infer from a system A to a system B, even if some of the 'stuff' is the same. Parker seizes on the notorious problem of inferring from drug tests with rats (A) to the use for humans (B). But does this complication make materiality less important, or does this render simulation a better justification? No! Translating A to B, from rats to humans, is even more problematic in simulation, where less 'of the relevant respects' are included and one actually may not know what the 'relevant respects' are. Translating A to B requires knowing all differences between A and B and their respective relevance to the problem at hand - something that in the end can only be achieved by material measurements. This becomes evident in drug testing: tests with rats are at most a first step to consider its application to humans, only after material tests on a sample of humans are successful is a drug certified for humans. No simulation is sufficient to replace the final authority of material tests. Thus, similarity can adopt a justificatory role only if it is justified by material measurements—but then, why is the "intense focus on materiality" "somewhat misplaced"?

There is a second argument Parker uses against the special role of materiality and higher epistemic power of experiments: since experiments are impossible, e.g., to predict "noontime temperatures in various cities" (p. 492), simulations are "more likely to provide desired information about a target system" (p. 494) than an experiment. We agree that it is virtually impossible to put a laboratory experiment "into an initial state that reflects enough of the real atmosphere's complicated temperature structure" (p. 492). Moreover the laboratory should extent over 100's of kms to include long-range effects on the weather dynamics. Yes, we agree with Parker: systems exist, where no meaningful (human) experiment can be performed. While it would be more convincing if Parker would analyze the justificatory role of simulation and experiment for target systems, where both can work, let us address Parker's example. First, we do not deny the high quality of simulations for weather forecast. Indeed for any prediction inter - and extrapolations and models (remember also the risks discussed in Sect. 8) are needed. Strictly speaking, no experiment can predict, since always different conditions apply in a real life environment. But do these limitations reduce the special role of materiality?

The question at stake is, if the causal connections at work have to be justified by material measurements or if they can be derived from simulations. For this one has to address how the model is derived and furthermore how one obtains the conditions to run a simulation. As to the first problem, the equations to forecast weather are all founded on and are extensively validated in material measurements. But it would be mistaken to assume that the current knowledge about the atmosphere and the interactions with the environments does not require additional active research. Indeed, a wide range of research based on material measurement attempts to improve our understanding of the weather system. Such improved understanding is impossible from simulations alone and indeed the atmosphere itself and the earthly environment serve as a system of natural experiments, observed with millions of material measurements provided by buoys, ships, aircraft, satellites, and earth bound stations all over the world. As to the second problem, even the best algorithms and models for weather forecast predict nothing by themselves, and cannot be put in the realistic 'initial state' of a simulation. To become similar with the target system, the models must be fed with distributions of temperatures, strengths and directions of air streams, etc., at an initial time t_0. But just running the simulation may fail: it is a common experience that predictions of noon - time temperatures are not stable, but are constantly updated. While the models and algorithms of the simulation remain the same, it is the updated material input that causes the predictions to change. In the terminology of Parker, one may say that the similarity of simulation and the material world is insufficient and can only be resurrected by material measurements. Inferences to the world using simulations can at most be as good as the material input and need to be updated with material measurements for times longer than the typical fluctuation time of the relevant parameters. These frequent checks and adjustments of a simulation to material measurements is what we had already seen in Sect. 5 for the trustworthy simulation of electrons.

Thus, also for systems, where direct human intervention is impossible, materiality provides the necessary foundation. Indeed, to restrict materiality to human experiments is too narrow a view, instead it should include natural experiments (cp. Sects. 3 and 8.3). In this sense, Parker's example only highlights the particular importance of materiality.

Similar arguments can be made for cases that Winsberg (2009) addresses. The simulationist that Winsberg tasks with understanding supersonic jets in black holes can provide inferences to the world only by, at minimum, claiming agreement with the results by the experimentalist dealing with the same problem using material measurements. As pointed out before, to use these results for Winsbergs's aim to understand black hole dynamics, requires additional material information and modelling, which would require a separate discussion.

In the end, Winsberg concludes that models of the solar system based on Newton's laws have a stronger epistemic power than experiments (p. 591, see Sect. 8.3). Again, if Winsberg simply wants to state that human experiments are impossible for some systems, we agree. Otherwise, again, Newton's equations by themselves do no provide any information about the position of planets, but they have to be fed with measurements at a certain time t_0, the precision from calculating the positions at a later time t_1 will be inferior to measurements. Furthermore, as we have discussed in Sect. 8.3, a simulation based on Newtonian gravitation would fail to describe the

solar system precisely - a failure that is remedied by general relativity. Is this a minor disturbance of Winsberg's argument? We do not think so, since it relates to the core of where simulations and material measurements differ. One may be confident of a model and assume one has a trustworthy simulation, but material measurements show their limitations, an argument already brought forward by Morgan (2005). Does this really award simulations a stronger epistemic power than material measurements, as Winsberg argues?

10 What follows for the epistemic hierarchy?

In the previous section we analyzed some claims on epistemic parity of simulations in literature. We argued against parity between simulation and material measurement, since all these claims were required to anchor simulations in material measurements in multiple ways. While simulations by themselves do have 'epistemic power', it implies that, along Morgan (2005) (p. 323) justifications of results from a simulation are significantly more involved than those obtained from experiments—simply speaking, in addition to material measurements, something has to come on top.

We will now turn to assessing epistemic hierarchy using general definitions of the status within a hierarchy: firstly, adopting the general notion of hierarchy, for example used in social systems, and secondly, credibility, the generally accepted measure of epistemic hierarchy (see Sect. 2.2).

10.1 The epistemic dependence of simulation on experiment

Throughout our discussion, we noticed an asymmetry between material measurements and CSs. In brief, CS depends on material measurements: material measurements are mandatory for CS. The other way round, simulations are not mandatory for material measurements, although, no doubt, CS helps to focus and optimize material measurements and it is a tool for derived measurements.

The dependence of CS on material measurements was an overarching theme for the seven conditions that a simulation can attain trustworthiness. In the words of Winsberg (2009), simulations grow out of a "long history of experiment and observation" (p. 591). The priority of material measurement in time is the most apparent and uncontroversial claim concerning the ranking of simulation and measurement.[16] But the dependence of simulations on material measurements is not only in time, but epistemic: measurements are mandatory to obtain knowledge from simulations. Without material measurement, a simulation cannot be validated, and material measurements provide the properties to start a simulation. Simulations are, at most, as precise as their material input. There is no reciprocal role for simulations.

What does such dependency mean for the respective ranking in a hierarchy? At least for social systems, according to Child (2019) (p. 1), "[h]ierarchy is a system in which

[16] Occasionally theory and thus simulation may precede observations, as possibly for Black Holes, supersymmetric particles etc.. However, they gain trustworthiness only after agreement with observations is established.

the members .... are ranked according to their status or authority," creating an unequal relationship and subordination. Indeed, simulation and material measurement have an unequal role in the system of science and the overarching dependency of simulation on materiality does not justify parity, but suggests a superior role of measurements.

10.2 Are experiments and simulations equally credible?

This conclusion is emphasized by the criterion of credibility (see Sect. 2.2). To make the case, assume a trustworthy simulation agrees with observations at a certain stage, but more precise or qualitatively new observations disagree with the simulation. Thus, either simulation or material measurement has to be adjusted, they cannot be on a par. Scientific practice is to change simulation and not the material measurement.[17] Among the numerous examples, let us pick out climate change, often cited as witness for the parity of simulation: one observes a sustained, faster than expected warming of oceans (Cheng, Lijing and others 2019). It is not the measurement that will change, but its results will be adopted in future simulations. Another example is from the LHC: many physicists expected the properties of the 125 GeV enhancement to look different from the SM. They were ready to discard the SM, rather than the measurement. It appears obvious that material experiments are given higher credibility, and thus higher epistemic status.

A litmus test to decide on the relative credibilities and epistemic statuses is thus: if the two concepts of material measurement and simulation lead to results that disagree, then that concept is epistemically superior, that scientific practice leaves unmodified while requiring the other one to be adjusted.

It is beyond this paper to discuss why a material measurement has a higher credibility than simulation, but for us, a natural reason, as many suggest, is its material connection with the target system. The higher credibility that at least scientists assign to material measurements compared to models and simulation also becomes apparent in the attempt to replace models in simulation as much as possible with data, or at least data–driven methods. It would be interesting to understand how the higher credibility is explained by those that argue that materiality is not particularly important.

11 Conclusion

By analyzing simulations at the forefront experiment at the LHC, we identified seven requirements that are jointly necessary for a simulation to be trustworthy and useable as a proxy for measurements without dedicated and repeated validation with data. These requirements are also applicable to simple models and simulations of remote systems. While the requirements are necessary, they are not sufficient. They are possibly met when fixed standards are repeatedly applied, as with many engineering problems, but when analyses address new property regions, simulation requires dedicated validation,

[17] Note that, as mentioned in Sect. 3, we only consider technical correct simulations and experiments. We do not deny that in case of a confounding material measurement, first the measuring device is (again) scrutinized and indeed a 'loose cable' might be identified. However, if at all, this is just a delay in the conclusions and epistemically not relevant in the long term.

and thus is not trustworthy anymore *per se*. Simulations of systems with strongly varying and fluctuating parameters are also likely to be untrustworthy if they are used to predict the systems's state at times longer than the typical fluctuation of relevant parameters.

While we agree with Winsberg (2009)(p. 591) that "the time has long passed' since 'we did not have sufficient systematic knowledge of nature" to build simulations, we also note that the scale of engineering and scientific problems, like those at subnuclear or cosmological distances, have also grown, often rendering simulations too imprecise. Simply said, if one is interested in the general state of a system, simple models are sufficient, while simulations including several additional effects are more precise. However, the most precise inference to the world requires continuous input by material measurements. If a model or a simulation is trustworthy depends on the scientific goal, notably the required precision.

We argued against the claim of epistemic parity between simulations and material measurements. Firstly, running CSs to infer to the world without material input is futile. (Trustworthy) simulation can only produce relevant results if material measurements provide reasonable properties for which they can start running and, at least in many cases, a continuous update of parameters. For many applications of trustworthy simulations it is a misconception to believe that the CS is just run by itself. Regular checks and updates of the simulation outputs by material measurements are a rather general procedure. In the common notion of hierarchy, such dependence of simulations on material measurements can be translated into a lower status. Secondly, assuming credibility as a measure for the hierarchical level, as is implicitly suggested in the literature, scientific practice assigns higher credence to material experiments: in case the result of a simulation disagrees with a material measurement, it is general scientific practice that the simulation be adjusted.

Naturally, such conclusions depend on what is considered material measurement and simulation. For example, downplaying and blurring the distinction between CS and traditional experiment, as Parker (2009)(p. 487) does, one easily finds comparable epistemic power. Similarly, if material measurements are restricted to (intervening) experiments, it seems incoherent to call for less focus on materiality using systems, where experiments are impossible. In contrast, we do not restrict measurements to human interventions, but include observations of natural experiments. Not the least, this reflects that, at the LHC, measurements are a hybrid of experimental practices and observation.

While we deny simulations are on a par within the epistemic hierarchy, we are far from the "undervaluing" of the "epistemic worth" simulations, that Morrison (2015)(p. 243) is afraid of. Simulations are a convenient tool to transform a direct measurement into a derived one. More importantly, models and simulations are necessary to understand the relations between results of different experimental measurements such that they can be embedded in a (more or less) general theoretical framework. Thus, simulations are an important tool at the LHC and other measurement facilities. However, their aim and status is significantly different from material information, and thus should be clearly distinguished.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Amsler, C., et al. (2008). Review of Particle Physics. *Physics Letters B, 667,* 1–1340.

Anderl, S. (2016). Astronomy and Astrophysics. In P. Humphreys, A. Chakravartty, M. Morrison, & A. Woody (Eds.), *The Oxford Handbook of Philosophy of Science* (pp. 652–670). Oxford University Press.

Bähr, M., and others. (2008). Herwig++ physics and manual. *The European Physical Journal C,58*(4), 639–707.

Barberousse, A., & Jebeile, J. (2019). How Do the Validations of Simulations and Experiments Compare. In C. Beisbart & N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (pp. 925–942). Springer.

Beisbart, C., & Saam, N. J. (2019). *Computer Simulation Validation - Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. Springer.

Boge, F. J., & Zeitnitz, C. (2020). Polycratic hierarchies and networks: what simulation-modeling at the LHC can teach us about the epistemology of simulation. *Synthese*

Chall, C., King, M., Mättig, P., & Stöltzner, M. (2019). From a Boson to the Standard Model Higgs: A Case Study in Confirmation and Model Dynamics. *Synthese,* 1–33.

Cheng, L. (2019). How fast are the oceans warming? *Science,363*(6423), 128–129.

Child, J. (2019). *Hierarchy*. Routledge.

Cranmer, K. (2015). Practical Statistics for the LHC. In *Proceedings, 2011 European School of High-Energy Physics (ESHEP 2011): Cheile Gradistei, Romania, September 7–20, 2011* (pp. 267–308).

Dorigo, T. (2016). *Anomaly!* Collider Physics and the Quest for New Phenomena at Fermilab: World Scientific.

Englert, F., & Brout, R. (1964). Broken symmetry and the mass of gauge vector mesons. *Physical Review Letters, 13,* 321–323.

Evans, L., & Bryant, P. (2008). LHC Machine. *Journal of Instrumentation, 3*(08), S08001–S08001.

Franklin, A. (2013). *Shifting Standards: Experiments in Particle Physics in the Twentieth Century*. University of Pittsburgh Press.

Giere, R. N. (2009). Is computer simulation changing the face of experimentation? *Philosophical Studies, 143*(1), 59–62.

Guala, F. (2002). Models, Simulations, and Experiments. In L. Magnani & N. J. Nersessian (Eds.), *Model-Based Reasoning: Science, Technology, Values* (pp. 59–74). Springer.

Gueguen, M. (2020). On robustness in cosmological simulations. *Philosophy of Science, 87*(5), 1197–1208.

Higgs, P. W. (1964). Broken Symmetries and the Masses of Gauge Bosons. *Physical Review Letters, 13,* 508–509.

Hoeche, S., Krauss, F., Schonherr, M., & Siegert, F. (2013). QCD matrix elements + parton showers. The NLO case. *Journal of High Energy Physics*, 2013(4).

Lenhard, J., & Hasse, H. (2017). Boon and bane: On the role of adjustable parameters in simulation models. In M. Carrier & J. Lenhard (Eds.), *Mathematics as a Tool*. Tracing New Roles of Mathematics in the Sciences: Springer Verlag.

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

Mättig, P. (2019). Validation of Particle Physics Simulation. In C. Beisbart & N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (pp. 631–660). Springer International Publishing.

Mättig, P. and Stöltzner, M. (2019). Model landscapes and event signatures in elementary particle physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*.

McNeile, C. (2010). High-Precision c and b Masses, and QCD Coupling from Current-Current Correlators in Lattice and Continuum QCD. *Phys. Rev. D,82*, 034512.

Morgan, M. S. (2005). Experiments versus models: New phenomena, inference and surprise. *Journal of Economic Methodology, 12*(2), 317–329.

Morrison, M. (2015). *Reconstructing Reality: Models, Mathematics, and Simulations*. Oxford University Press.

Murray-Smith, D. J. (2019). Verification and Validation Principles from a Systems Perspective. In C. Beisbart & N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (pp. 99–118). Springer International Publishing.

Oberkampf, W. L. (2019). Simulation Accuracy, Uncertainty, and Predictive Capability: A Physical Sciences Perspective. In C. Beisbart & N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (pp. 69–97). Springer International Publishing.

Parke, E. C. (2014). Experiments, Simulations, and Epistemic Privilege. *Philosophy of Science, 81*(4), 516–536.

Parker, W. S. (2009). Does matter really matter? Computer simulations, experiments, and materiality. *Synthese, 169*(3), 483–496.

Parker, W. S. (2015a). Computer Simulation, Measurement, and Data Assimilation. *The British Journal for the Philosophy of Science, 68*(1), 273–304.

Parker, W. S. (2015b). Getting (even more) serious about similarity. *Biology & Philosophy, 30*(2), 267–276.

Peschard, I. (2011). Is simulation a substitute for experimentation?

Roush, S. (2018). The epistemic superiority of experiment to simulation. *Synthese, 195*(11), 4883–4906.

Rumsfeld, D. (2002). News briefing on february 12, 2002. Department of Defense.

Schael, S., et al. (2006). Precision electroweak measurements on the Z resonance. *Physics Reports, 427,* 257–454.

Somerville, R. S., & Davé, R. (2015). Physical Models of Galaxy Formation in a Cosmological Framework. *Annual Review of Astronomy and Astrophysics, 53,* 51–113.

Staley, K. W. (2018). Securing the Empirical Value of Measurement Results. *The British Journal for the Philosophy of Science, 71*(1), 87–113.

Tal, E. (2017). Measurement in Science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.

The ATLAS Collaboration. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B, 716,* 1–29.

The ATLAS Collaboration. (2014). Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data. *European Physical Journal C: Particles and Fields, 74*(10), 3071.

The ATLAS Collaboration (2016). Search for resonances in diphoton events at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Journal of High Energy Physics*, 2016(9).

The ATLAS Collaboration. (2018a). Measurement of the Higgs boson coupling properties in the $H \rightarrow ZZ^* \rightarrow 4\ell$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP, 03,* 095.

The ATLAS Collaboration. (2018b). Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D, 98,* 052005.

The ATLAS Collaboration. (2019). Electron and photon performance measurements with the ATLAS detector using the 2015–2017 LHC proton-proton collision data. *JINST, 14,* P12006.

The ATLAS Collaboration. (2019). Measurement of the four-lepton invariant mass spectrum in 13 TeV proton-proton collisions with the ATLAS detector. *JHEP, 04,* 048.

The ATLAS Collaboration (2019). Measurements of the Higgs boson inclusive, differential and production cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$TeV with the ATLAS detector. Technical report.

The CMS Collaboration. (2012). Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Physics Letters B, 716,* 30–61.

The CMS Collaboration. (2016). Search for Resonant Production of High-Mass Photon Pairs in Proton-Proton Collisions at $\sqrt{s}$ =8 and 13 TeV. *Physical Review Letters, 117*(5), 051802.

The OPERA Collaboration. (2012). Measurement of the neutrino velocity with the OPERA detector in the CNGS beam. *JHEP, 10,* 093.

Trucano, T. G., Pilch, M., & Oberkampf, W. L. (2002). General Concepts for Experimental Validation of ASCI Code Applications. *SAND,* 2002–0341.
Winsberg, E. (2009). A tale of two methods. *Synthese, 169*(3), 575–592.
Winsberg, E. (2010). *Science in the Age of Computer Simulation*. The University of Chicago Press.
Winsberg, E. (2019). Computer Simulations in Science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition.
Woodward, J. (2003). Experimentation, Causal Inference, and Instrumental Realism. In *The Philosophy Of Scientific Experimentation* (pp. 87–118). University of Pittsburgh Press.