Minds and Machines (2022) 32:43–75
https://doi.org/10.1007/s11023-021-09569-4

SI: MACHINE LEARNING: PREDICTION WITHOUT EXPLANATION

Two Dimensions of Opacity and the Deep Learning Predicament

Florian J. Boge[1]

Received: 1 December 2020 / Accepted: 1 August 2021 / Published online: 3 September 2021

Abstract
Deep neural networks (DNNs) have become increasingly successful in applications from biology to cosmology to social science. Trained DNNs, moreover, correspond to models that ideally allow the prediction of new phenomena. Building in part on the literature on 'eXplainable AI' (XAI), I here argue that these models are instrumental in a sense that makes them non-explanatory, and that their automated generation is opaque in a unique way. This combination implies the possibility of an unprecedented gap between discovery and explanation: When unsupervised models are successfully used in exploratory contexts, scientists face a whole new challenge in forming the concepts required for understanding underlying mechanisms.

Keywords Machine learning · Opacity · Models · Explanation · Scientific understanding · Exploratory experimentation

1 Introduction

Deep neural networks (DNNs) are currently being praised for their astonishing utility in applications ranging "from biology to cosmology to social science" (Jordan & Mitchell, 2015, p. 255). Popular-level accounts even go so far as to proclaim a "revolution in scientific research" (Royal Society and Alan Turing Institute, 2019, p. 1), or that DNNs are "changing the way we do science".[1] At the same time, DNNs are notoriously associated with the label 'black box', which is usually meant to say that a DNN corresponds to "a function that is too complicated for any human to comprehend" (Rudin, 2019, p. 206).

Given this black box-nature, how can DNNs *truly* help us advance science? For, in the words of Raghu and Schmidt (2020, p. 27; original emphasis):

[1] See https://dx.doi.org/10.1126/science.aan7049.

✉ Florian J. Boge
fjboge@uni-wuppertal.de

[1] Interdisciplinary Centre for Science and Technology Studies (IZWT), Wuppertal University, Gaußstr. 20, Room S.11.19, 42119 Wuppertal, Germany

> Many standard applications of deep learning [...] focus on *prediction*—learning to output specific target values given an input. Scientific applications, on the other hand, are often focused on *understanding*—identifying underlying mechanisms giving rise to observed patterns in the data.

Given the close connection between understanding and explanation (de Regt, 2017; Grimm, 2010; Khalifa, 2017; Rudin, 2019; Strevens, 2008), scientists' interest in what is usually called 'eXplainable AI' (XAI) should be rather high accordingly. But under one reading at least, "'explanation' here refers to an understanding of how a model works, as opposed to an explanation of how the world works" (Rudin, 2019, p. 206). Hence, even if XAI succeeds, can we really expect an understanding of "underlying mechanisms" or "how the world works" from it?[2]

In this paper, I will offer a nuanced response to this question, by arguing for the following three theses: (i) Deep learning (DL) models, appropriately construed, are *instrumental* in a specific sense that sets them apart from many (though not all) traditional scientific models, including computer simulations (CSs). (ii) XAI concerns two distinct kinds of black box-ness, or *opacity*, and reducing one will not necessarily aid in reducing the other. These may be seen as two *dimensions* to the opacity-problem in DL—a notion I shall make precise below. (iii) This unique combination of opacity and instrumentality implies that we cannot generally expect to understand the mechanisms underlying (decisive patterns in) the data when these are successfully recognized and predicted by DL algorithms.[3] In particular, when certain conditions are jointly met, it is highly likely that DL allows new discoveries that scientist will have a hard time understanding.

The main goal of this paper is hence to make sense, from a philosophy of science point of view, of claims to DL revolutionizing or changing science. Establishing (i)–(iii) requires some conceptual effort though. First, I will distinguish three senses in which DNNs are models (Sect. 2.1), and distinguish the sense appropriate for my purposes. Subsequently (Sects. 2.3 and 2.4), I will then determine the relevant sense of instrumentality, and why it makes DL models *non-explanatory*.

Section 3 will define the notion, and argue for the existence, of two dimensions to opacity in DL; the implication being that DNNs are opaque in a way that is *not reducible* to the (well-known) opacity of CSs. In Sect. 4, I will then show how DNNs' instrumentality and opacity together can lead to unprecedented gaps between discovery and explanation. That, together with DL's unprecedented success in handling big data, I call the *DL predicament*.

[2] I will presuppose a kind of pluralism about 'explanation' here: Given, for instance, the quantum nature of the physics examples discussed below, *causal* explanation is probably not the right concept. But this is clearly different for the other, biological case study.

[3] 'Mechanism' should be construed rather broadly here. For instance, the Higgs mechanism defies various features typical of mechanisms (Lyre, 2008; Smeenk, 2006), but for most physicists still counts as sequence of steps that promotes an understanding of the underlying physics.

2 Deep Learning Models

2.1 Three Senses of 'Model' in Deep Learning

In the DL literature, the use of 'model' and 'representation' abounds, but instances of DL are often equally referred to as 'algorithms' or simply 'techniques'. This is rarely accompanied by an explication; something that has raised philosophers' attention before (Humphreys, 2013; Napoletani et al., 2011).

Napoletani et al. (2011, p. 13) actually refrain from calling DNNs 'models' altogether and solely use 'technique'. Humphreys (2013, p. 580), on the other hand, acknowledges the possibility of "simulating neural dynamics" with DNNs, but also urges to "keep separate uses of neural nets as simulation models from their use as techniques in computational science", and additionally finds most neural nets to be "extremely crude models of real brains[...]."

The latter verdict is frequent in the literature (e.g. Chirimuuta, 2020; Goodfellow et al., 2016; Sullivan, 2019), not least because feed-forward processing and gradient descent are biologically implausible; although notions such as 'distributed representation' or 'representation learning' suggest a stronger connection. *Spiking* neural nets could be more promising in this respect (e.g. Kasabov, 2019), and interest in the brain-DNN correspondence persists. However, assuming that we take the simulation of brain dynamics as the *relevant* sense in which DNNs can be models, we would be able to understand only biological brain processes by means of them.

There is a further notion of model applicable to the DL context:

> Using [...] data we build a prediction model,[...] which will enable us to predict the outcome for new unseen objects. (Hastie et al., 2013, p. 2)

> Fundamentally, Machine Learning is using algorithms to extract information from raw data and represent it in some type of model. We use this model to infer things about other data we have not yet modeled. (Patterson & Gibson, 2017, p. 1; emphasis omitted)

> The goal of modeling is to develop a parametrized mapping between the data domain and the response set. [...] In machine learning, the modeling, itself, may have several algorithms to derive a model; however, the term algorithm here refers to a *learning algorithm*. (Suthaharan, 2016, p. 7; original emphasis)

This notion of model as some parametrized input–output mapping is closely related to the universal approximation theorem, which in essence says that a DNN "can approximate virtually any function of interest to any desired degree of accuracy, provided sufficiently many hidden units are available" (Hornik et al., 1989). Newer theorems (Poggio et al., 2020) also show that this can be accomplished with gradient descent and in finite times.

Thus, a *trained* DNN may be considered a model in the sense of an input–output mapping that characterizes patterns in the data, ideally capable of accurately predicting further points to that pattern, or even a new phenomenon. For instance, obtaining a single point in the output space could amount to the recognition that a

bunch of data indeed classify as indicative of a new token of some type of phenomenon of interest. An example of this kind is astrophysicists' recent discovery of four new pulsars with the aid of a DNN called 'SPINN' (Morello et al., 2014).

However, the relation between output and new phenomena can also be more indirect: If the task is statistical and the output is a label for classification, the distribution of data points into classes can reveal an unexpected excess of data that fall into a certain class. In turn, this might indicate a so far *unknown* phenomenon, responsible for the data-excess. This situation obtains, for instance, in particle physics, as shall be discussed in more detail below.

Note that the *learning algorithm* involved in deriving this model may itself count as yet another model: Buckner (2018) points to the possibility of understanding *concept abstraction* on the basis of deep convolutional nets, without drawing too close a parallel to either brain processes or most details of human cognition. Similarly, it may be possible to understand certain errors made by DNNs in analogy to errors made by humans under similar conditions (Buckner, 2021, for discussion). But the analogy between human and machine learning can only be taken so far; for instance, it remains an open question "whether current or future DNN architectures can implement compositional recursive grammar" (ibid, pp. 4–5). Thus, what happens during the training of DNNs can also serve as an *abstract* model of *aspects* of human learning, independently of the brain-machine (or a close mind-machine) analogy.

In sum, at least three distinct senses of 'model' should be distinguished here, which, so far as I can see, exhaust the use of 'model' in the DL literature:

(a) DNNs as (crude) models of actual brains,
(b) the algorithms employed in DL as abstract, selective models of human learning, and
(c) the input–output mappings approximated through training as models of features pertaining to the data, such as their statistical distribution.

Which of these, if any, is the sense relevant for understanding underlying mechanisms? As pointed out above, if we took (a) to be the relevant sense, our understanding would be limited to brain processes. The same applies to (b) and human learning. However, sense (c) is fairly general, and hence does not share these problematic features.

2.2 Prediction, Discovery, Explanation

It is exactly this mapping, established during the training phase, that provides DNNs with their *predictive* capabilities. To see this, recall that "[n]early all of deep learning is powered by [...] stochastic gradient descent" (Goodfellow et al., 2016, p. 147), which means the iterative minimization of a 'loss function' through several rounds of training (also called 'epochs'). Hence, during the training, DNNs are forced to do better and better at some kind of task, regardless of whether given access to

class-labels (supervised learning)[4] or clustering the data without such guidance (unsupervised learning).

But the training stage thus also amounts to an iterative fitting of the model to a training set: It proceeds by a successive change of free parameters in response to the 'loss' experienced when offering a certain output for the data points encountered. If this is done carefully so as to avoid over- and underfitting to the training set (and with some tricks such as unsupervised pre-training of individual layers),[5] a DNN can excel in handling so far unencountered examples.

Now once the training is over, the DNN's parameters are fixed and the model in sense (c) is established. But letting a trained DNN loose on actual data of interest, it may be able to exploit the patterns encountered during training—of which the scientist may be fully unaware—to successfully predict further points to that pattern. As we saw above, this ideally leads to the recognition of new, scientifically interesting phenomena such as pulsars, i.e., to new *discoveries*.[6]

Discovery, when connected to a theory, model, or method, is clearly intimately linked to that theory's, model's, or method's predictions. For instance, Lakatos (1970, p. 116) makes an identification between a theory's (verified) "excess empirical content over its predecessor (or rival)" and "the discovery of novel facts." Similarly, Maher (1988, p. 282; emph. added) argues that "successful prediction provides reason to think that a *discovery method* is reliable".

Generally speaking, neural networks are capable of providing predictions in the *strong* sense of forecasting the occurrence of a novel, previously unobserved phenomenon—which sometimes is to be construed in the more general sense of 'use-novelty' (Worrall, 1985), here meaning that information about that phenomenon was not included in training and model-definition.

For instance, a shallow network for language processing was recently able to forecast the discovery of novel thermoelectric materials from the textual content of scientific papers (Tshitoyan et al., 2019). When benchmarked on historical papers published before a certain date, between some 20–45% of the network's top 50 predicted materials had been discovered with a span of some 3–18 years past that respective date (ibid., p. 97). Similarly, a combination of an unsupervised algorithm (*k*-means) for clustering spatio-temporal climate data into characteristic patterns with a (supervised) convolutional neural net was recently used to forecast the occurrence of certain weather patterns 5 days ahead, with an accuracy of some 90% (Chattopadhyay et al., 2020).

[4] Supervised techniques comprise both classification and regression tasks. Given, however, that there is a close connection between both, and that the latter can sometimes even be treated as the continuum limit of the former (cf. Skansi, 2018, p. 61; fn. 14), I will not be too careful in distinguishing them in this paper.

[5] Following a suggestion by an anonymous referee, I should note here that these 'tricks' by themselves can already decrease understanding. For instance, choosing the *learning rate*, i.e., the hyperparameter scaling the gradient in stochastic gradient descent, can not only drastically influence the training speed but also determine whether the training gets stuck. This latter effect is, however, poorly understood (Goodfellow et al., 2016, p. 417).

[6] Caveat: I will always mean 'discovery' in the sense of 'discovery of empirical phenomena', not as in the theoretical discovery of some sort of mechanism or scientific hypothesis.

However, we can gather from the above quote by Hastie et al. (2013)[7] that most 'predictions' made by (D)NNs are certainly to be understood in a considerably *weaker* sense, namely as predicting a certain data point (or a set thereof) to fall under a certain class (or to be attached a certain value more generally), which, ideally, corresponds to the recognition of the presence of a type of phenomenon of interest (as with the pulsars discovered by SPINN). Moreover, for any prediction, strong or weak, to count as *successful*, and to thus provide a new discovery, it has to be confirmed by (further) empirical means; as was (obviously) the case with the thermoelectric materials and weather patterns, but also the candidate pulsars (cf. Morello et al., 2014, p. 1659).

Now following Douglas (2009, p. 458), we can hold predictive accuracy to also be a key marker of scientific *explanations*:

> A scientific explanation will be expected to produce new, generally successful predictions. An explanation that is not in fact used to generate predictions, or whose predictions quickly and obviously fail, would be scientifically suspect.

Accordingly, I take it that in inquiring about underlying mechanisms, we are inquiring about an *explanatory model* that matches the successful predictions of a DNN.[8] However, it will thus be the burden of this paper to first argue that the DL model in sense (c) (which shall always be meant by 'the' DL model below) is itself not explanatory.

2.3 Instrumentality of Deep Learning Models

What does a DL model actually represent? To approach this question, consider the simple, shallow (i.e., single-hidden-layer) network in Fig. 1. Its two input nodes may be collectively represented by the vector $\boldsymbol{x} = (x_1, x_2)^t$. Similarly, $\boldsymbol{h} = (h_1, h_2)^t$ corresponds to the hidden layer, and may here be assumed to compute a function $\boldsymbol{h} = \mathbf{max}\{W\boldsymbol{x} + \boldsymbol{b}, \mathbf{0}\}$, where the vector-valued 'max' applies component-wise, W is a matrix of weights, and $\boldsymbol{b}$ a bias vector.

The edges leading from the input to the hidden layer in the diagram may be understood as transmitting the input with a certain weight, and the nodes as setting the received values off by a bias. The non-linear 'max'-function corresponds to a given node's activation upon receipt of the (weighted, biased) input. This repeats at the edges from $\boldsymbol{h}$ to the output-layer y, albeit with a weight-*vector* $\boldsymbol{w}$, a single bias c, and no non-linearity. For a given set of weights and biases, the network in Fig. 1 thus computes the function $y(\boldsymbol{x}) = \boldsymbol{w} \cdot \mathbf{max}\{W\boldsymbol{x} + \boldsymbol{b}, \mathbf{0}\} + c$.

[7] See Goodfellow et al. (2016, pp. 98–102), Skansi (2018, pp. 51–56), Suthaharan (2016, pp. 130–141) for further evidence that this is a common use of 'prediction' in machine learning circles.

[8] I find myself in good company with this verdict: Guidotti et al. (2018, p. 12), for instance, define what they call the 'black box explanation problem' as "providing a global explanation of the black box model through an interpretable and transparent model [that] should be both able to mimic the behavior of the black box and [...] should also be understandable by humans."

For a more complex network, there would be several hidden layers, vectors would usually be longer, activations could be different non-linearities, and the output vectorial. But the general description would not change: The network would still correspond to a function $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{y} \circ \boldsymbol{h}^{(n)} \circ \cdots \circ \boldsymbol{h}^{(1)}(\boldsymbol{x})$.

Assuming now that $x_i \in \{0, 1\}$, weights and biases can be changed by a learning algorithm such that $y(\boldsymbol{x})$ spits out 1 exactly if one input is 1 and the other is 0, and 0 otherwise.[9] It thus provides a model of the *exclusive or* (Goodfellow et al., 2016, Sect. 6.1). However, even *interpreting* $y(\boldsymbol{x})$ in this fashion requires prior *conceptualization* of the data in terms of truth values.

Features of ambiguity in scientific models are certainly nothing new: Any mathematical model, when conceived of purely as a formal structure, admits of multiple interpretations. But for traditional models, the interpretation is achieved by assigning meanings to its mathematical symbols. Now as was pointed out above, we could interpret weights, biases and activations in terms of properties of axons and neurons, or maybe the relevance associated to the entries of $\boldsymbol{x}$ by a learner; but that would lead us to interpreted models in sense (a) or (b), not (c).

To see the difference more clearly, turn to Fig. 2. The mathematical model depicted in (a) is, famously, that of a damped harmonic oscillator. In the first instance, the variable X represents the height of the oscillation, t a parameter (such as time) along which the oscillatory pattern spreads, the derivative(s) to the changes of the oscillation over time (or the changes of these changes, respectively), γ (indirectly) to the damping, and ω to the frequency of oscillations (and so indirectly to the period).

However, multiple real world systems (approximately) realize these relations: X could represent the position of a mass attached to a spring, and then γ would, for instance, represent the friction that applies to spring and mass. But X could also represent, say, such things as the concentration of a chemical in a sample (for $X = 0$ a baseline concentration), (a smoothing of) the number of individuals in a population (with similar meaning for $X = 0$), and various other things. Furthermore, the parameter t need not even represent time, but could just as well be (say) some non-linear function thereof. With all these changes in interpretation, the meaning of the model as a whole would be changed.

Now the meaning of the output y of the shallow network depicted in Fig. 2b will certainly covary with the interpretation of $\boldsymbol{x}$ as either, say, simple propositions to be combined into complex ones, activations of diodes in a circuit, or the more general (non-)occurrences of two mutually exclusive events. But the same is not true about the non-linearity $\boldsymbol{h}$ or the parameters $\boldsymbol{\theta} = (W, \boldsymbol{b}, \boldsymbol{w}, c)$. For a DL model, weights and biases are merely parameters to be adjusted automatically during training and $\boldsymbol{h}$ represents a hyperparameter to be chosen in advance. Beyond that, they are in general not assigned any specific meaning at all: Just consider how for a deep network with some hundreds or even thousands of nodes, no scientist will presumably be able—*or even bother*—to sort out what each and every weight and bias might represent.

[9] Notably, this is possible only with at least one hidden layer, because the function to be learned is not linearly separable (Buckner, 2019; Goodfellow et al., 2016; Minsky & Papert, 1969).

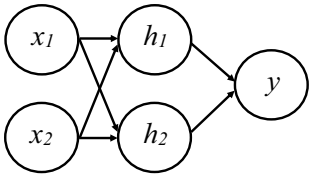

Fig. 1 Shallow neural network capable of learning the exclusive or

Accordingly, it is not immediately clear what (hyper)parameters should be taken to represent about the curve in Fig. 2b, and certainly even less so as to what they represent about the system whose behavior is in turn represented by that curve. Given this apparent meaninglessness of the parameters, I submit that, in contrast to traditional mathematical models, the interpretation of a DL model stems entirely from the *conceptualization of input and output*. Without this conceptualization, prior to the training stage, we would not be able to recognize the meaning of a DNN's predictions at all.

An anonymous referee has confronted me with an interesting objection in this connection. First of all, none of the above implies that a DNN's (hyper)parameters *cannot* be assigned a meaning at all: In DNNs used for image recognition, for instance, specific nodes can be associated, via their activations, with specific features of the images the network operates on, such as edges or hues. Secondly, a large chunk of the technical literature is devoted to making such associated features explicit, and so it seems possible to think that a DNN has *itself* developed an internal model of patterns in the data at the end of the training, which model could be quite directly understanding-conveying.

I appreciate the point. However, I should here make explicit the stance on models I employ, which is strictly speaking incompatible with the foregoing assessment. Like, e.g., Potochnik (2017), I view models not as disembodied platonic entities, but rather, as epistemic devices used by cognizing, *conscious* agents. Furthermore, I am highly skeptical of the notion that (at least at present) DNNs can be literally viewed as such agents. Hence, unless explicitly constructed by the scientists *using* the DNN, that internal model is literally nowhere, or rather, does not really exist. Put differently, talk of a model internal to the DNN must on my view be seen as a metaphor for the fact that interpreting the DNN in the right ways can suggest a way forward to new models that can promote understanding.

A comparison to traditional *statistical* models also suggests itself at this point, at least when these are used in a fully data-driven way, and not generated on the basis of a conceptual model or background theory of the entity or process under study. Statistical models may be generally identified with parametrized distributions, and of course their parameters are (under the aforementioned circumstances at least) equally used to fit a given model to data. The major commonality between traditional statistics and DL is that, in general, the meaning of these parameters has to do rather with properties of the data than with the underlying mechanisms generating them.

However, the parameters of statistical distributions usually at least have clear meanings as, say, the most probable value or a characteristic width, whereas it is

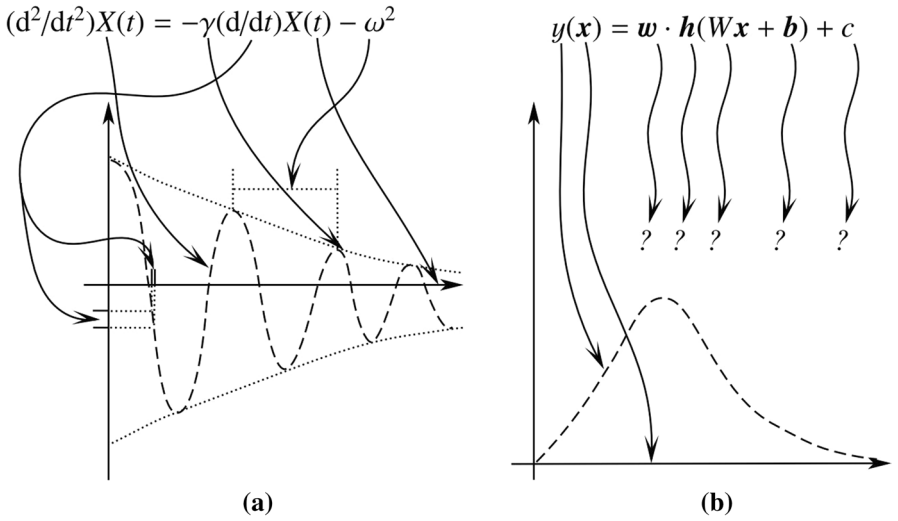

Fig. 2 Differences between the interpretation of classical mathematical models (**a**) and DL models (**b**)

unclear that DL parameters have a representational function at all, as we saw above. Usually (though maybe not always) traditional statistics and DL also operate at different levels of *generality*[10] and are employed for different *purposes*:

> statistical methods have a long-standing focus on *inference*, which is achieved through the creation and fitting of a *project-specific* probability model. [...] By contrast, DL concentrates on *prediction* by using *general-purpose* learning algorithms to find patterns in often rich and unwieldy data (Bzdok et al., 2018, p. 233; emph. added).

The choice of a class of DL models from which to choose, i.e., the general type of parameterized function to be adapted during the training, is effected by means of *hyperparameters*: those parameters, like number of layers, nodes, or even choice of activations, determining the general *architecture* and those, like learning rate or batch size, determining the *training process*.[11] The difference in generality attested by Bzdok et al. (2018) thus amounts to the fact that a DNN-architecture and its training are usually chosen with a general *kind of task* in mind, not the detailed properties of a single, concrete data set to be evaluated.

In contrast, for classical statistical models, the choice is usually dictated by far narrower concerns:

[10] There are no 'free lunches' though (Wolpert & Macready, 1997), and present DL also still falls short of providing anything like 'general intelligence' in the sense of a domain-general ability to recognize and exploit patterns at a human-like level (e.g. Lyre, 2020, for a recent assessment).

[11] See, for instance, https://towardsdatascience.com/what-are-hyperparameters-and-how-to-tune-the-hyperparameters-in-a-deep-neural-network-d0604917584a for an excellent overview.

> The choice of a class of models, gamma, lognormal, Weibull, is initially based on the *shape of the data*[...]. The question arises as how this decision can be explicitly incorporated in the analysis. A Bayesian 'solution' would be to include all 'reasonable' models and then to base the decision on the relative likelihoods. A frequentist approach would be to perform a goodness-of-fit test for several possible models. (Davies, 2014, 44; emph. added)

The choice of a goodness-of-fit metric must be based on the properties of both the data as well as the considered distributions (e.g., continuous vs. discrete). Similarly, likelihood-based methods can choose among different information criteria (AIC, BIC, etc.) that mediate in different ways between fit and simplicity (Sober, 2002). However, choosing the 'reasonable' statistical models, as well as the criteria for selecting among them, thus already requires a respectable amount of insight into the *structure* of the data.

As pointed out above, fixing a class of DL models by means of hyperparameters in contrast rather requires empirical knowledge about performance in a certain kind of task, as well as considerations of speed and the ability to generalize from the training set. Only then is the chosen architecture used to "*find* patterns in often rich and unwieldy data" (Bzdok et al., 2018, p. 233; emph. added). In addition, the choice of an eventual model is effected by an automated learning algorithm—an all but trivial point when it comes to DNNs' ability to promote an understanding of the data-generating mechanisms, as we shall see.

Finally, consider another class of traditional, tunable models with relatively low conceptual content, sometimes called 'phenomenological' in the philosophical literature (e.g. Bokulich, 2011; Craver, 2006). To give an example: The Rydberg formula, $\lambda^{-1} = R(n_1^{-2} - n_2^{-2})$ (n_1, n_2 positive integers with $n_2 > n_1$), 'saves the phenomena' (spectral lines), but unlike Bohr's atom model is not rich enough in theoretical elements to offer an explanation of their occurrence (Bokulich, 2011, p. 41 ff.; Massimi, 2005, p. 36; Wilholt, 2005, p. 155).

According to Bokulich (2011, p. 44), phenomenological models are "only of instrumental value to the scientist", often "constructed via an ad hoc fitting of the model to the empirical data", and "useful for making predictions, but [...] do not purport to give us any genuine insight into the way the world is." Most of these things appear to be true of DL models. However, 'phenomenological' has too many different meanings (Suárez & Cartwright, 2008, p. 70) and there is also no clear, prior association between DL models and 'phenomena'. In that respect, DL models are certainly closer to what Harris (2003, p. 1510) calls 'data models'. However, via the output variable $\boldsymbol{y}$, the functions instantiated by DNNs add information that certainly transcends a mere cleaning and interpolation of data points. Hence, they are also not data models in Harris' sense.

To properly classify DL models, I hence suggest to focus on a particular *aspect* recognized by Bokulich, and call them *instrumental*. As we saw above, the specific *sense* of instrumentality here is that of being *instrumental-qua-devoid of content*—call that 'c-instrumental': Most elements in formal representations of a DL model need not be assigned any meaning at all in order for the model to have predictive value. This notion I take to subsume, next to DL models, also

phenomneological models (at least on some reading of the term), as well as statistical models (at least when not derived from some conceptually rich theory).

2.4 Instrumentality and Understanding

The intended sense of instrumentality here is not to be conflated with another sense prominent in the models-debate. This other sense refers to the employment of *unrealistic assumptions* (see Basso et al., 2017, p. 424). Call that 'r-instrumental'. However, whether the employment of such assumptions cannot lead to *more* than predictive and instrumental value remains a controversial issue (ibid.).

An often discussed example is Schelling's model of segregation (Schelling, 1971), in which the housing and moving behavior of two different kinds of agents (e.g.: blacks and whites) is modeled by (e.g.) black and white dots on a chess board-like arrangement with filled and vacant fields.

As Schelling demonstrated, even modest preferences of model-agents for equally-colored neighbors suffice to produce segregation patterns. However, in the 1970s it was not clear whether the model could be adequately linked to empirical evidence, and it actually builds on *various* unrealistic assumptions (e.g. Reutlinger et al., 2018, pp. 1076–1077). Accordingly, the model's status is controversial: Does it deliver an explanation of how actual segregation patterns arise, or a mere 'how-possibly' explanation?

In any case, it is agreed upon that the model provides *some* sense of understanding, by demonstrating that institutional racism does not *have* to be assumed in order to explain segregation (Grüne-Yanoff, 2013, pp. 855–856). On top of that, the distinction between how-possibly and potential how-actually explanations has been put into question by Bokulich (2014, p. 335), who shows that how-possibly explanations sometimes correspond to how-actually explanations when several details are abstracted away. Finally, today there is some amount of evidence in support of a mechanism relevantly *similar* to Schelling's (Card et al., 2008; Clark, 1991; McCrea, 2009).

The point, then, is that, regardless of its many unrealistic assumptions, Schelling's model *does* offer a (possible, or potentially actual) explanation of segregation patterns, and so arguably provides understanding.

A demarcation against c-instrumental DL models trades on a particular understanding of 'understanding', and it is instructive to consider several details of de Regt's (2017, p. 31 ff.) celebrated account of understanding in this connection. De Regt takes understanding on the basis of models to be possible if they are explanatory *qua representational* (also Giere, 2006, Chapter 4). A similarly important role for representation is reserved by Morrison (1999, p. 63):

> The reason that models are explanatory is that in representing [their target] systems they exhibit certain kinds of structural dependencies. The model shows us how particular bits of the system are integrated and fit together in such a way that the system's behaviour can be explained.

Hence, establishing ways in which to represent a certain target by means of a model allows us to map the relations established in the model onto relations pertaining, for all we know, to the target, and so, if the model's behavior matches that of the target in relevant respects (e.g., segregation patterns emerge), we may infer an explanation of the observed target-behavior from the model (e.g., in terms of moving behavior being in part determined by preferences for neighborhood-composition).

However, note that data themselves are usually assumed to have "some sort of representational content", and their "curation and classification [...] involves *interpretative* decisions" (Leonelli, 2019, pp. 4, 11; emph. added). Hence, does the relation established by a DL model not equally represent something about the underlying mechanisms?

This is, in a way, certainly correct. But it merely points us to the fact that representation is not all when it comes to explanation and understanding. In de Regt's account, for representational models to explain, they must also be constructed under the principles of an *intelligible theory*, where a theory is intelligible if it has certain qualities that "provide *conceptual tools* for achieving understanding" (de Regt, 2017, p. 118; emph. added).[12] Among these tools, de Regt (2017, p. 115) lists "visualization, mathematical abstraction, and causality [as] prime examples." An example for the use of mathematical abstraction is the development of an "intuitive feeling for how quantum-mechanical systems in two-slit-like situations behave, by familiarity with the linear character of the Schrödinger equation." (ibid., p. 113) Hence, it does not (necessarily) concern straightforward deductive use of mathematics, but (in general) rather heuristic qualitative reasoning with mathematical concepts.

Clearly, replacing 'understanding' by a notion such as 'intelligibility' or 'grasping' (Strevens, 2013) looks like replacing one unanalyzed, primitive notion with another one which is just as opaque. However, de Regt spells out the intelligibility of a theory in terms of "qualities [...] that facilitate the *use* of the theory" (de Regt, 2017, p. 40; emph. added), and Reutlinger et al. (2018, pp. 1084–1085) equally offer a use-oriented, empirically accessible explication of Strevens' notion of grasping.

It is not my aim to reconstruct or defend these accounts in further detail here. Rather, I take away the general lesson that understanding the mechanisms in a targeted domain on the basis of models requires that these models contain representations that are *conceptually rich enough* to make those mechanisms intelligible.

Assuming a notion of understanding along these lines, it is straightforward to see why a trained DNN is instrumental in ways that can impair understanding. Just recall the discussion following Fig. 2: The elements of the model that could be used as representations are weights, biases, and activations. But for the sake of (c), these are merely adjustable parameters devoid of content, not representations that help us conceptualize some underlying mechanisms by facilitating visualization, qualitative

[12] The focus on theories may not do full justice to understanding from models: Morrison famously emphasizes their (partial) autonomy. Nevertheless, Morrison and Morgan (1999, p. 31) also hold a "process of interpreting, conceptualising and integrating" that goes in during model *construction* ultimately responsible for understanding.

mathematical reasoning, or causal inference.[13] Hence, c-instrumentality, in contrast to r-instrumentality, *directly* threatens scientific understanding.

Without a doubt, it is *possible* to understand something about underlying mechanisms on the basis of the outputs of a DNN though. Hence, what is required to facilitate that understanding? I believe that the following three steps are crucial in that respect: (I) The *conceptualization* of input and output, prior to training; (II) establishing what the (trained) DL model represents, on account of (I); and (III) *connecting* that represented something to underlying mechanisms on the basis of further background knowledge.

An example might be helpful here. For reasons to become clear, *particle physics* provides an excellent case study for my purposes, so I will turn to it several times. Beside the fact that particle physics has been an "early proving ground" (Cho, 2017) for machine learning in general, physicists at CERN face more than 200 petabyte of stored data from the LHC in their analyses—'big data science' indeed.

A typical problem is the definition of *likelihood-ratios* for hypothesis testing, because the theory is intractable and the high dimensionality of the feature space (multiple particles with energies, momenta, charges, etc.) does not allow the generation of enough simulated data for approximation (Baldi et al., 2014).

A DNN can reduce dimensionality by replacing data with a class identity ('signal' or 'background'). Classification is a 'supervised' task, meaning that a DNN learns based on *labelled* data. If labels are defined by parameterized *cuts* in the feature space (e.g.: energy above a certain threshold), whose optimum position is then learned by performing stochastic gradient descent, the DNN can approximate a likelihood-ratio by providing a proportion of signal to background data. If the meaning of the cuts is well-understood, moreover, the distribution of data across the signal/background-divide can then reveal whether it is somewhat justified to assume that a sought for particle has been produced.[14]

To see how (I)–(III) apply here, first note that there are several 'levels' of data in particle physics: *raw* data strictly speaking correspond to the "byte-stream of the readout from detector electronics" (Albertsson et al., 2018, p. 19). However, these are usually interpreted immediately as indicating the energy deposited by a particle in a specific component of th detector while traversing it. Data referred to as *reconstructed* in particle physics correspond to tracks gathered from these isolated interactions, and already obtain a vastly richer interpretation: The shape of these tracks across different layers of the detector allows the attribution of features such as particle type, momentum transverse, and angles relative, to the beam direction, or even 'missing' energy physically expected but not measured (signifying a neutrino). Finally, certain *higher-level* features, defined as usually non-linear functions from the (interpreted) reconstructed data, can be used as data for DL algorithms as well.[15]

[13] For evidence that this is in line with standard views of much of the XAI community, just consider the quote in Fn. 8 again.

[14] For a detailed treatment, see Voss (2013).

[15] However, even the lowest level conceptualization in terms of energies deposited in detector components at a given time need not in any way be submitted to the algorithm in order for it to perform

As can be seen, several conceptual steps are involved in data-preparation, prior to subjecting them to analysis. Similarly, slicing the input-space in such a way that some data count as 'signal-like', i.e., typical of decay chains containing the relevant particle, some as 'background-like', involves a conceptual step. Actually, however, even the *data-taking* already involves conceptual steps. As Harris (2003, p. 1512) has famously pointed out, many "instruments do not produce uninterpreted [...] data", and this is in a way also true of LHC-detectors: A sophisticated *trigger system* is required to select manageable amounts but the selection criteria installed in the three trigger-levels are based on *physics hypotheses* (Karaca, 2018, Sect. 4). Altogether, this covers step (I).

As pointed out above, Baldi et al. (2014) used DNNs to approximate likelihood-ratios. Hence step (II) consists in training a DNN with appropriately labeled data in such a way that it can be interpreted as approximating this ratio. But that would not be possible had the data not been conceptualized in terms of physics variables and meaningful cuts.

Finally, step (III) means interpreting an excess of signal over background data in terms of the sought for particle. Clearly, this is only possible based on (II), together with the fact that likelihoods are probabilities conditioned on relevant hypotheses. For illustration, see Fig. 3. Note also the involvement of further explanatory models emphasized therein.

As we saw already in the simple network of Fig. 1, the interpretation of the function approximated covaries with the conceptualization of the input.[16] A suitable learning algorithm could always approximate *some* given input–output mapping. But it would be impossible to tell what that mapping represents if the meaning of the data was left unspecified. Similarly, for being able to *explain* the predicted outputs in terms of, say, connections, building blocks, and currents in a circuit, we would first need to recognize that the DL model predicts activations that match an XOR gate.

I take it that this analysis in terms of steps (I)–(III) is an assessment of understanding from DL, or the possibility of a want thereof, compatible with Sullivan's recent appraisal of what she calls 'link uncertainty', i.e., the "lack of scientific and empirical evidence supporting the link that connects the model to the target

Footnote 15 (continued)
accurately: advances in image recognition suggest that it may be possible to harvest successes using raw detector data directly (Albertsson et al., 2018, p. 8).

[16] An anonymous referee has pointed out to me that, in the words of LeCun et al. (2015, p. 439), "Deep neural networks exploit the property that many natural signals are compositional hierarchies, in which higher level features are obtained by composing lower level ones. In images, local combinations of edges form motifs, motifs assemble into parts, and parts form objects." I appreciate the point, and aspects concerning the structure of the data and their processing by DNNs will become important later when I consider questions of opacity. However, for now note that when a data vector is fed to a (convolutional) DNN, this data vector only represents an image insofar as we interpret it to refer to colors and hues distributed across a 2D pixel grid. Furthermore, if we reinterpret both the input and output vector to (say) an autoencoder as representing (say) the properties of atoms distributed across a crystal lattice, that would immediately also change what $y(\boldsymbol{x})$ represents: In the first case, it would represent the salient, or 'crucial' features of the image (as specified by the reconstruction loss-function); in the second case, it would rather represent an emergent property of the crystal, which is only visible as soon as most of the detail is abstracted away (so far as compatible with that same loss function).

phenomenon" (Sullivan, 2019, p. 8). Because of the dependence on (I) and, especially, (III), the DL model *on its own* is conceptually too poor to provide an understanding of underlying mechanisms: Only if, via (I) and (II), a connection can be made to those mechanisms in a final step, (III), will a DL-success promote an understanding of them.

3 Two Dimensions of Opacity

3.1 Opacity: From Computer Simulations to Deep Learning

Much of what has been said about r-instrumental models straightforwardly extends to CSs. A common analysis has the generation of a CS start from what Morrison (2015) calls a *conceptual model*. Such a model then (usually) needs to be discretized, and translated into computer code, to serve as a proper basis for the steps undergone by the computer.

Analyses of the modeling steps involved in devising a CS of this general type are now fairly common, and found, in varying detail and explicitness, in e.g. Winsberg (1999, 2010), Humphreys (2004), Beisbart (2012), Hasse and Lenhard (2017), Boge (2019b), or Durán (2020).

However, several modifications to this general scheme have been suggested in the literature as well: Hasse and Lenhard (2017), for instance, amend the sequence by an explicit acknowledgement of loops in the modeling process. That does not impair the general type of account though, for it simply means the iteration of several steps in the modeling chain, stimulated by comparison between simulated and observed data.

Some observations by Lenhard and Winsberg (2010) are more troubling: In complex simulations, as used in climate science, models become highly entrenched. Hence, a linear sequence which facilitates explanations does not seem applicable. Nevertheless, Boge and Zeitnitz (2020) identify (close to) linear *sub*structures in a similarly complex simulation environment, and argue that these remain applicable as a descriptive account of initial modeling steps.

The point is *not* that this immediately restores a straightforward path to explanation and understanding from CSs. Depending on the complex relations between these substructures, as well as the meaning and use of various parameters in each of them (see Hasse & Lenhard, 2017), it can still be very hard to infer anything even remotely explanatory from the results of a CS.

The upshot rather is that none of these modifications impairs the observation that the initial modeling step in a CS consists in a conceptualization *of the target*: If the target is a complex system such as the earth's global climate or a scattering event at the Large Hadron Collider, it may be necessary to conceptualize parts of the target individually first and then connect them up—which can, over an iterative process of mutual readjustment, lead to entrenchment and take scientists away from the initial meaning of individual models. But the first steps in designing a CS are nevertheless very different from the first steps in designing a DL model.

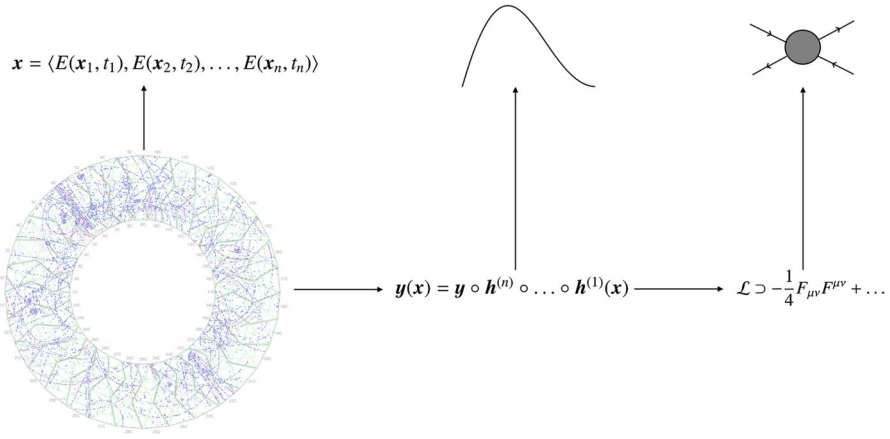

Fig. 3 Illustration of steps (I)–(III) in the particle case, here illustrated with raw instead of reconstructed data (see fn. 15). Horizontal arrows indicate (formal) modeling steps; vertical arrows indicate conceptual/interpretive steps. Raw data image taken from https://cds.cern.ch/record/1409759/files/event67hires.png (©CERN 2011)

Now it is also clear that designing a CS often requires additional assumptions that may introduce (further) artefacts. In the words of Lenhard (2019, p. 224):

> Simulation [...] works with relatively 'weak' objects [...] that are, to an important extent, codefined by modeling decision, adjustments, and discretizations.

Codefinition by modeling choices is nothing special in CSs, as was pointed out already above (also Frigg & Reiss, 2009, p. 598 ff.). Similarly, discretization is crucial for CSs, but numerical techniques were introduced long before the first CS was run. But some *adjustments* are certainly special in CSs: A prominent example is the so-called 'Arakawa operator' (Lenhard, 2007; Winsberg 2010), which was introduced to fix the instability of the atmosphere in a global climate simulation of the 1950s, on the pains of introducing the deliberately unphysical assumption of conserved kinetic energy (known to also dissipate as heat). The reason for this instability was, however, not mathematical but computational: Digital computers can only handle finite-place approximations to decimal numbers.

This example underscores that, next to entrenchment-problems, unrealistic assumptions can become so dominant that a CS might loose explanatory value. The difference to DL really is that this situation can at least *in principle* be improved upon by (i) tracking the relations between partial models in detail, (ii) using more realistic assumptions, when improved hard- and software allow this, (iii) using a less distorting discretization method under the same conditions, and so forth. Whenever

this is possible and the amount of r-instrumentality can be (assessed and) contained in this way, scientists become able to infer explanations *directly* from CSs.[17]

The difference between CS and DL may be summarized as follows: The former begins with a conceptualization of the *target*, and from that predicts 'hypothetical data'. The latter begins with a conceptualization of *data*, and from that *equally* predicts new, hypothetical data. A deeper connection to the target needs to be established *post hoc* in DL, whereas it is made (or at least attempted) *ante hoc* in CS. Arguably, this carries over to a major difference in explanatory potential.

However, there is also a major *commonality*: Both CSs and DL have been recognized for their *opacity* (cf., in particular, Burrell, 2016; Humphreys, 2009). Following Humphreys' seminal definition, I take this to mean the following:

> a process is epistemically opaque relative to a cognitive agent X at time t just in case X does not know at t all of the epistemically relevant elements of the process. (Humphreys, 2009, p. 618)

I take it for granted that epistemic opacity is relative to an agent and involves a lack of knowledge. The process in both cases is the computational process, and its opacity is usually traced back to the *complexity* of the algorithm (cf. Burrell, 2016, p. 5; Humphreys, 2009, p. 619).

Now in DL, this is partly also conditioned on a "mismatch between mathematical optimization in high-dimensionality characteristic of Deep Learning and the demands of human-scale reasoning and styles of semantic interpretation" (Burrell, 2016, p. 2). This is underscored, for instance, by the existence of 'adversarial' examples in image recognition, wherein a small, dedicated perturbation of an image, which is imperceptible to humans, can lead to a radical misidentification (Goodfellow et al., 2014). To some extent, these can be explained by considering the 'learning context' and the limitations imposed by the finite classification made available to the network, but there are certainly also many features that remain puzzling (e.g. Buckner, 2020, 2021, for discussion).

The point hence is that, despite some abstract analogies between human and Deep Learning, it is in important respects opaque *how* the machine learns. Call that *h-opacity*. H-opacity concerns the way in which a DNN automatically alters the instantiated function in response to data. However, by that token it effectively just adds to *complexity-related* opacity:[18]

> Though a Deep Learning algorithm can be implemented simply in such a way that its logic is almost fully comprehensible, in practice, such an instance is

[17] E.g. Boge (2019a) or Durán (2017, 2020) for recent accounts of how this is possible. Note that no specific notion of explanation is presupposed here; Boge (2019a) discusses examples wherein a deductive-nomological and a functional explanation are being inferred from CSs, respectively.

[18] In fact, depending on the agent in question, h-opacity may concern all three forms of transparency in complex computational systems identified by Creel (2020), i.e., "functional transparency, or knowledge of the algorithmic functioning of the whole[;] structural transparency, or knowledge of how the algorithm was realized [...]; and [...] run transparency, or knowledge of the program as it was actually run in a particular instance, including the hardware and input data used [...]." (ibid., p. 569) I take it, however, that *functional* transparency is the main target of XAI, w.r.t. h-opacity.

> unlikely to be particularly useful. Deep Learning models that prove useful [...] possess a degree of unavoidable complexity. (Burrell, 2016, p. 5)

Since h-opacity thus concerns the complexity associated with the algorithm (including the learning-prescription), it is *continuous* with the opacity of CS. But it is unclear whether, or to what extent, this sort of opacity impairs understanding:

> In order to gain understanding of [...] mechanisms of segregation, one does not need to know whether Schelling's model was implemented using a functional, object-oriented, or actor-based language[...]. More drastically, [...] one does not even need to know whether the model was implemented on a computer system at all or whether it was implemented on a checkerboard[...]. Thus, implementation back-boxing in itself does not undermine our ability to explain or understand phenomena. (Sullivan, 2019, pp. 12–13)

This assessment resonates well with various proposals on the opacity of CSs. For instance, Durán (2018, p. 108; emph. added) argues that "researchers are only interested in a limited amount of information that counts for the *justification* of results."

For Durán, this allows disputing the *epistemic relevance* of the unknown elements, and so whether CSs are even interestingly opaque at all. Similarly (Lenhard, 2006, pp. 611–613), who *embraces* CSs' opacity and considers it "a major obstacle to explanatory potential" (Lenhard, 2019, p. 224), still acknowledges that CSs promote understanding:

> a researcher can acquire a kind of orientation within the model [...] based on experience of the model's behavior [....] mediated by the calculating machine[...], whereas the model itself remains epistemically opaque. (Lenhard, 2006, p. 613)

Regardless of which side we take in this debate in detail, the tenor which is common to all these positions clearly carries over to DL's h-opacity: In order to gain understanding of underlying mechanisms from DL, we need not understand the training or the learning algorithm in full detail (Sullivan, 2019). However, whatever potentially *prevents* understanding in DL, in the sense of a disconnect with underlying, data-generating mechanisms, must therefore be something else.

3.2 What Was Learned?

H-opacity is one sense in which DL can be a black box, but does it exhaust DL's black box-nature? Consider the following assessment by Raghu and Schmidt (2020, p. 27):

> Interpretability methods are sometimes equated to a fully understandable, step-by-step explanation of the model's decision process. [...] Instead, research in interpretability focuses on a much broader suite of techniques that can provide insights ranging from (rough) *feature attributions*—determining what input features matter the most, to *model inspection*—determining what causes certain neurons in the network to fire.

Notice that 'interpretability' is used synonymous with 'explainability' here. In contrast to the quote by Rudin (2019) from the introduction, however, we see that more may be at stake with'explanability' than merely understanding how the machine works.

Actually, Raghu and Schmidt acknowledge a "rough split in the type of interpretability method", according to whether it focuses on model interpretation or feature attribution. In a very similar vein, I shall here argue that there are two independent *dimensions* to the opacity-problem in DL, of which h-opacity is one, and which only roughly coincide with this (rough) split.

The second dimension concerns the question of *what was learned* by the machine. Call that *w-opacity*. As I shall show, w-opacity is, ultimately, the distinctive factor which sets DL apart from *all* traditional models and, eventually, impairs our ability to acquire scientific understanding in a special way.

In the next section, I will offer a criterion for the existence of two dimensions, and demonstrate how it applies in practice. The purpose is to show that w-opacity is *non-reducible* to h-opacity (which is continuous with CSs' opacity), and so that there is a novel challenge. Subsequently, I will argue that this unique combination of c-instrumentality and w-opacity is likely to lead to an unprecedented gap between scientific discovery and understanding, at least when DL is used under certain conditions of interest in several sciences.

First, however, I should make precise the sense of *opacity* here. For that purpose, recall the four central elements of Humphreys' definition: a process, an agent, unknown elements of the process, and the epistemic relevance of the unknowns for the agent.

The agent could at present be essentially any member of the scientific community: Even computer scientist are mostly aware that DNNs are remarkable, but neither generally understand their functioning nor what it is about given scientific data that drives their success. This is why XAI is such a hot topic.

In fact, the unknowns are what makes w-opacity (and DL, accordingly) special: They correspond to *automatically discovered insights*; complex, non-obvious features that can be abstracted from the data and allow the machine to discriminate. Their existence is an empirical matter, so I will provide examples below.

It is these very features that drive predictive success but, as the examples will show, at the same time yield the greatest prospects for understanding. They are hence epistemically relevant.

What may be least obvious is the process involved. It would be tempting to refer to the underlying mechanisms themselves, for they of course *generate* those non-obvious features. However, that would conflate steps (II) and (III): In a sense, it is always 'opaque' what gives rise to novel data, until (or unless) explanatory models are available. That has nothing to do with DL per se.[19]

[19] Additionally, even data-production involves a human component (Hacking, 1992), as do storing, preparation, and dissemination (Leonelli, 2016). Hence, the features in question could be artefacts of data-generation and handling, or stem from an ill-chosen *format* for the purposes at hand. Before exploring underlying mechanisms, these and similar issues need to be sorted out.

As a matter of fact, it is easy to recognize the *very same* process involved in h-opacity as involved also in w-opacity. This is what it *means* that there are two dimensions to 'the' opacity-problem in DL, instead of two problems. When a DNN learns to approximate a desired function, it is hence not only opaque how, precisely, it achieves this goal: It is also opaque what it is about the data that drives this process.

For illustration, we may return to the physics case study. Baldi et al. (2014) actually performed a benchmark, aimed at estimating the potential of DNNs to discover new physics. DNNs here significantly outperformed shallow networks and boosted decision trees on well-understood, simulated data. However, by how much the performance differed was highly dependent on the *kind* of input.

As was noted above, physicists distinguish between 'low-level' and 'high-level' features: The former are more or less directly inferred from (the distribution of) electrical signals in the detector, the latter constructed as (usually non-linear) functions of the former. An example of a low-level feature, to recall, is the momentum-component transverse to the beam pipe associated with a particle track, and one of a high-level feature is the reconstructed (invariant) mass of a particle that decayed before interacting with the detector.

The surprising result of Baldi et al. (2014) was that the DNN always outperformed the other algorithms when given access only to the low-level features, and had a modest additional increase when given access also to the high-level features. The other algorithms instead exhibited major differences in performance between these situations. From this, Baldi et al. (2014, p. 7; emph. added) concluded "that [DNNs] are *automatically discovering* the insight contained in the high-level features."

These automatically discovered high-level features are a clear instance of the unknown 'whats', but their existence is by no means restricted to particle physics. In the life sciences, for example,[20] DNNs have recently excelled in predicting protein structures from amino acid data, in the form of distances between amino acid residues. The researchers here also inquired "how the network arrives at its distance predictions", hoping to further "understanding of the folding mechanisms" (Senior et al., 2020, p. 714).

Don't be misguided by the 'how' though: though: Senior et al. used integrated gradients to map out "*input features* that affect the network's predictions" (Senior et al., 2020, p. 714; emph. added), and from this concluded that "the network is using *intermediate predictions* to discover important interactions and channelling information from related residues" (ibid.; emph. added). For instance, for pairs of residues in direct contact, "all of the highest attribution pairs are pairs within or between the secondary structure that one or both of the output pair(s) are members of." (ibid.) Hence, attribution maps suggested that the DNN exploits

[20] A similar point is also made by López-Rubio (2020) about visual categories in convolutional and generative-adversarial networks used for image-recognition and production respectively. Note also that López-Rubio (2020, p. 1; emp. added) is careful to describe the corresponding states as being "*interpreted by humans as* complex visual categories".

information on protein *sub-structures*, somehow contained (but not plainly visible) in amino acid-data.

Now, certainly, information on the location of residues with secondary protein structure is vastly more informative regarding the production of the protein that corresponds to the spatially ordered amino acids in terms of a folding mechanism than the mere statement of that spatial information. Similarly, the information that a particle with given mass must have been produced as an intermediate state in a decay chain is vastly more informative regarding the production of particle tracks in terms of an elementary scattering process on the sub-nuclear level than the mere statement of those tracks. This illustrates quite vividly why the complex features learned by DNNs, but hidden from plain sight, should be considered epistemically relevant.

These examples establish the sense in which DNNs are not just h- but also w-opaque—something that connects more closely to questions of understanding 'the world' rather than 'the machine'. 'But', you may insist, 'is it not equally opaque what features of initializations to a *CS* drive *its* success?' I believe this to be a confusion: Due to the interplay between target-conceptualization and coding, all information about what makes initial values *play out* in terms of specific simulation outputs is contained in the algorithm, not the data used for initialization. As far as I can tell, there really is no pendant for w-opacity in CSs or other scientific models.

3.3 Independence

As was pointed out above, recognising h- and w-opacity as two dimensions means showing the non-reducibility of the latter to the former. For dimensions usually characterize *independent* features: The dimension of a vector space, say, corresponds to the maximum number of linearly independent vectors in that space.

This captures the relevant intuition, but is rather uninformative for showing the dimensions' existence. For that, we need a *criterion*, like the following:

(C_0) Two features of opacity to some process shall be considered independent in case they can be *addressed* independently.

The intuitive appeal of (C_0) may be seen in terms of a mathematical metaphor: If we consider the total opacity O_p of some process p to be a function of two variables, $O_p(h, w)$, then we can see their independence if we are able to keep one fixed while investigating changes of O_p under variations of the other. As we shall see below, something quite similar actually happens in certain studies on DNNs' opacity.

However, (C_0) is too unspecific to be used in practice. For that purpose, I suggest to pay attention to the *means* by which opacity is addressed. Following the above discussion, these means correspond to variables either characterizing the DL method (weights, biases, choice of activation, etc.) or the DL task (higher-level features, protein sub-structures, visual categories, etc.). Hence, refine (C_0) as follows:

(C_1) Two features of opacity to some process shall be considered independent in case they can be addressed by means of *disjoint sets of variables* that make reference to *distinct features* of the process, respectively.

Hence, if the opacity of process p can be addressed by variables that make reference to one set of features of p, and equally by whole other variables that make reference to a completely different set of features of p, I take it that this means addressing different dimensions of p's opacity in each case.

Using (C_1), it is sufficient to prove the existence of explainability-methods utilizing variables that refer to features relevant solely for addressing either kind of opacity respectively, as shall be done below. First note, however, that I do not claim that all or even most XAI-studies can be sorted according to the h/w-distinction. In fact, many studies address both dimensions at once, even if in unequal proportion. In this sense, there is indeed only a "rough split in the type of interpretability method".[21] But that doesn't impair *talk* of two dimensions: In terms of the earlier mathematical metaphor, this is just like saying that for many p from the class of DL algorithms, most methods reach points of low $O_p(h, w)$ by climbing down a path that reduces both h and w (Fig. 4 for illustration).

To see the existence of the dimensions now, first consider the study by Schwartz-Ziv and Tishby (2017). The authors used an information-theoretic framework to address the fact that "there is still no comprehensive understanding of the optimization process or the internal organization of DNNs" (Schwartz-Ziv & Tishby, 2017, p. 1). What Schwartz-Ziv and Tishby (2017) did was map out the paths followed by hidden layers in what they called the 'information plane', i.e., the plane defined by treating the mutual information $I(\boldsymbol{h}^{(n)};\boldsymbol{x})$ and $I(\boldsymbol{h}^{(n)};\boldsymbol{y})$ between hidden layers $\boldsymbol{h}^{(n)}$ and input $\boldsymbol{x}$ or *targeted* output $\boldsymbol{y}$ in a supervised task as axes of a Cartesian coordinate system.

A central result was that (tested) DNNs go through two phases in which they develop a representation that is in a sense informationally optimal. In the first phase, the network increases the information layers have on the desired output; in the second phase, information on the input is reduced, so as to remove 'redundancies' (cf. Schwartz-Ziv & Tishby, 2017, p. 3).

The details are not terribly important; neither is the fact that the validity of this result is "critically influenced by the nonlinearities employed by the network" (Saxe et al., 2019, p. 14), or only shown to hold for a certain range of connected tasks. What is important is that, firstly, the study contributed to an understanding of *how* certain DNNs learn to achieve successful performance.

Secondly, the *nature* of input and desired output was irrelevant to seeing two phases: Schwartz-Ziv and Tishby (2017) used dots distributed across a sphere, which could stand in for various real-world patterns. More precisely, these patterns were used as representatives for an equivalence class of tasks related by invertible

[21] Saliency maps, e.g., approximate the weights adjusted by the network during training to map out salient features of the input (cf. Simonyan et al., 2013). To get a handle on w-opacity, it is thus sometimes even useful to reduce h-opacity first.

transformations (Schwartz-Ziv & Tishby, 2017, p. 4). $\boldsymbol{x}$ and $\boldsymbol{y}$ thus figured as fixed parameters to the study, not variables. On the other hand, changing mutual informations characterise the dynamics of hidden layers—not the data.

Now contrast this with another study from particle physics, entitled *What is the Machine Learning?* (Chang et al., 2018). Chang et al. considered an extension of the Standard Model (SM) by a new boson that couples to SM-particles. Two distinct versions of the boson were considered: one coupling to right-handed and left-handed SM-particles identically and the other exclusively to left-handed ones.

Again, the physics-details are not terribly important; the only thing that matters is that the different coupling strengths predict different angular distributions of measurable particles, which can be assessed in terms of a quantity called the rapidity difference.

More interesting are the details of the method employed by Chang et al. What they did was weight all statistical distributions for a range of measurable quantities point-wise by the inverse height of the distribution of some particular, chosen variable at that point, thus removing all the information on that quantity from these distributions and flattening its own distribution into a uniform one. This they called 'data planing', in analogy to the smoothing of a surface in woodwork. In consequence, performance-drops were studied.

Besides the rapidity difference, also the reconstructed invariant mass of the particle was planed for. In the symmetric case (coupling to left- and right-handed particles identically), planing away the mass-information already obliterated the DNN's ability to discriminate. In the asymmetric case (coupling to left-handed particles only), the DNN remained at least somewhat able to discriminate signal from background data, so long as *only* the mass was planed for. When both rapidity difference and mass were planed for, however, predictions amounted to guesswork in this case as well.

This result concurs with the fact that DNNs rely on higher-order information that they abstract from the data autonomously. However, crucial for the question of two dimensions is that, firstly, the physical quantities planed away are obvious instances of (potentially unknown) 'whats'.

Secondly, for the sake of seeing which physical quantities yielded the discrimination-power, it was irrelevant how the network evolved during training. These facts figured as fixed parameters of the study. *Physics* variables characterizing the data, together with (global) performance-measures, conveyed insight into the reasons for *success*.

Now recall that (C_1) urges us to pay attention to the means (more precisely: variables) by which a given feature of opacity is addressed. Given that Schwartz-Ziv and Tishby (2017) used some arbitrary input from a broad equivalence class and a dummy-output to be achieved (i.e., some randomly chosen binary classification of patterns), the amount of information contained in a DNN's hidden layers on these during any given training epoch was used exclusively to characterize (the dynamics undergone by) the architecture. In contrast, neither physics variables nor global performance measures make any reference to that architecture at all. I hence take it that these two studies demonstrate convincingly that there are aspects of opacity in DL for which (C_1) is fulfilled. In other words: W-opacity constitutes an independent

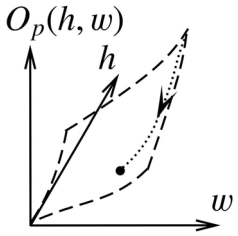

Fig. 4 Illustration of the metaphor explaining the difference between dimensions of opacity and the rough split between explainability methods. Dashed lines symbolize a surface of values O_p takes on over the $h - w$ quarter-plane. The dotted line indicates a path traced out by an explainability method

dimension to the opacity of DL, and may therefore create an independent epistemic problem (as I believe it does).

4 The Deep Learning Predicament

4.1 The Utility of Unsupervised Learning

I have argued that DL models are c-instrumental and opaque in a unique sense, but it is not clear yet what follows from this for scientific understanding. As we saw in Sect. 2.4, if steps (I)–(III) are executed skilfully, DL *will* promote understanding. In this section, I will finally demonstrate that (I)–(III) cannot always be realistically executed.

An important first step for seeing this is to realize that in the above examples it was straightforwardly possible to recognize the hidden, complex features exploited by DNNs (the unknown 'whats') *because these were benchmark-studies*. In the protein case, structures had been determined experimentally beforehand; in both particle studies, the data came from CSs, and so were well-understood in terms of conceptual physics models. In these studies, researchers were hence in possession of rich information about the targeted objects and used that information to tackles the question of what the network had learned.

Now the situation is clearly different when the goal is genuine *discovery*: As was pointed out in Sect. 2.2, DNNs can make predictions as to the occurrence or observation of future or so far unobserved events (strong prediction) or of a certain data point (or a set thereof) qualifying as indicative of a certain type of phenomenon (weak prediction). However, if, unlike in the pulsar example, the phenomenon is predicted to be of a *novel* type, it may not just be unclear what features of the data give rise to the prediction: It may even be unclear how to *identify* these.

Before turning to a realistic scenario in which this can happen, let me first point out that *unsupervised* learning plays a special role in this connection. The contrast

between supervised and unsupervised DL roughly corresponds to *classification*[22] versus *clustering* (Suthaharan, 2016, pp. 7–8): In a supervised task, training-data are *labelled*, and so the model learns to sort new data into pre-defined classes. Unsupervised models simply group together data based on structures encountered during training.

The usefulness of either depends on the kind of application: "Some domains, such as natural language processing, are known to benefit tremendously from unsupervised learning techniques" (Goodfellow et al., 2016, p. 412), and this is likely also true of searches for new phenomena in particle physics.

Deep autoencoders have in fact been trained in a *weakly* supervised fashion on particle data (cf. Farina et al., 2020), meaning that they learn to recognize only data labeled 'background' (i.e., 'not of interest'), and everything else is lumped into a catch-all 'anomaly'-class. It is thus possible to discover traces of new particles without relying on predictions from theory (which may not be available at all).

However, a major drawback to this approach is "the reliance on accurate background-only samples for training" (Farina et al., 2020, p. 6). It would be natural to employ CSs in producing those samples, but "[t]his would work only insofar as the [CS] accurately represents the background in the data[...][and] artifacts special to the [CS] are not learned by the autoencoder." (ibid.)[23]

Surprisingly, in an *un*supervised benchmark by Farina et al., "the autoencoder still succeeds in detecting anomalies in the test set even though they are present in the training set[...] as long as [it] does not see 'too many' anomalies in the course of its training[...]" (Farina et al., 2020, p. 7). Hence, unsupervised DNNs trained on supposedly well-understood data might able to recognize rare but poorly understood events. Moreover, because they are not subject to the same theoretical bias as (weakly) supervised models, they have a greater *discovery-potential* in that respect.

4.2 Unsupervised Exploration: Discovery Without Understanding?

Now consider the conditions under which remarkable discoveries are quite often made: *Exploratory* phases, which are not strongly guided by theory. Exploratory experimentation has been recognized for being special in many ways; most importantly in that "[t]he typical context of exploratory experimentation is the *formation* of [...] *conceptual frameworks*." (Steinle, 1997, p. S71; emph. added)

As an example, Steinle (1997, p. S72) discusses Faraday's introduction of the concept of magnetic force-lines, which ultimately lead to Maxwell's electrodynamics.[24] As we now know, this was an important step towards major scientific progress. But it took "two [...] decades" until even Faraday's "concept was fully developed" (ibid.).

It now comes in handy that we have chosen particle physics as a major case study, for it *is* about to enter an exploratory phase: Despite its predictive success, the

[22] ... and *regression*; cf., however, Fn. 4.

[23] Cf. Farina et al. (2020, p. 6) for similar problems with more data-driven approaches.

[24] See, however, Steinle (2016, Chapt. 7), for a wealth of further examples.

SM neither includes gravity nor dark matter, cannot explain neutrino oscillations, and has suspicious 'fine-tuning' properties. At the same time, many of the SM's favoured extensions have been ruled out by evidence, and future theoretical developments are far from obvious. Accordingly, the preamble of the 2020 update of the European strategy for particle physics emphasizes "the exploration of a new realm of energies" (European Strategy Group, 2020, p. 5).

Now recall, however, how physicists struggled greatly to make sense of the 'particle zoo' discovered in the 20th century. The discovery of muons, for instance, was famously greeted with the query "Who ordered that?" by Isidor Isaac Rabi. It is hence perfectly conceivable that the upcoming exploratory phase will reveal subtle traces of further unexpected particles. But now the following question arises: if such a discovery was powered by an unsupervised DNN, would physicists be able to make sense of it?

Given everything that was said so far, I believe that this is far from clear, and that highly similar problems may arise in other data-heavy sciences. For instance, consider the following verdicts from the earth science community:

> Unsupervised learning may aid the discovery of novel relationships[...] across the different dimensions of climate modelling [...]. A subsequent challenge for the Earth System community would be where an unsupervised approach reveals new system connections, requiring mechanistic understanding. (Huntingford et al., 2019, p. 5)

> [D]eep learning will soon be the leading method for classifying and predicting space-time structures in the geosciences. More challenging is to gain understanding in addition to optimal prediction, and to achieve models that have maximally learned from data, while still taking into account physical and biological knowledge. (Reichstein et al., 2019, p. 200)

To see where precisely the problems originate, first recall the difference between unsupervised and (semi-)supervised learning discussed in the previous section. In terms of underlying mechanisms, this difference plays out as follows: In a discovery based on supervised DNNs, the *labels* stem from a conceptualization of the target. Ideally, this will allow researchers to bypass w-opacity, because steps (II) and (III) will be fixed by some explanatory models' suggesting those labels.

Weakly supervised DNNs rely only on 'negative' labels, unsupervised ones on no labels at all. Thus, in both cases, the connection to underlying mechanisms becomes *severed*. In the unsupervised case, however, these observations extend even to models explaining the *non-anomalous* data. Thus unsupervised models not only yield the greatest discovery potential, but at the same time also the greatest disconnect to prior knowledge of data-generating mechanisms.

Now, if an unsupervised DNN (weakly)[25] predicts the presence of a novel phenomenon, physicists would certainly have several aces up their sleeves. They might

[25] Of course, it cannot be excluded that a DNN analyzing, say, the overall pattern of how known particles distribute across the existing data will be able to predict further particles at higher energies, i.e., to predict particles in the strong sense. However, at present, this seems to be mere speculation.

try to match it to existing proposals for new physics by adjusting certain free parameters in the corresponding physics-models. Or they might reassess their understanding of the background-physics in the domain where a significant anomaly is being indicated, with the goal of seeing whether the prediction was spurious.

However, absent any plausible physics model *or* reason for doubting the DNN's prediction, a real problem for explanation and understanding would arise from such an event. To see this clearly, consider also the importance of *background theories* in exploration (e.g. Franklin, 2005). The distinction Franklin makes between 'background' and 'local' theories in biology parallels similar distinctions made by Wallace (2020) and Karaca (2013) in physics. In detail, background theories determine the general structure of mechanisms in biology, or the structure of state spaces in physics, whereas local theories determine only the concrete ingredients to a particular mechanism or the state space for the problem at hand. In exploration, moreover, background theories "direct inquirers to the kinds of properties that could possibly have a [...] role in their local investigations[...]" (Franklin, 2005, p. 891).

Now at the very inception of present-day particle physics stand Rutherford's scattering-experiments (Duncan & Janssen, 2019, pp. 154), and famously, the first to make sense of several experimental findings here was Bohr. His explanation was based upon the assumption of quantized orbits for electrons (Duncan & Janssen, 2019, pp. 11–12), a development which contributed greatly to his later development of the atom model. Ultimately, this lead to modern-day quantum theory—a *then-new* background theory, which even introduced mathematics unfamiliar to physicists at the time.

The conceptual shift between classical and quantum theories is thus as radical as any (just think 'superposition' and 'entanglement'), which underscores Steinle's observations. Against the manifold successes of what is now called 'classical physics', such a leap must have clearly seemed inconceivable to many at the time (just recall Kelvin's infamous 'two clouds'). Yet it happened, stimulated by empirical findings that could not be properly conceptualized within the classical framework.

The point, then, is this: While the role Franklin ascribes to background theories in exploratory research may be correct in principle, exploration can even induce the need for *new* background theories. Given also the profound surprises particle physicists have faced in the past, we can't exclude big conceptual shifts lurking at unexplored energies.

A little more precisely, particle physics' present background theory is quantum field theory (QFT), so most candidate physics explanations would presumably be presented in terms of some QFT-Lagrangian. However, given, e.g., the well-known difficulties of integrating Einstein's general theory of relativity into QFT, it is by no means certain that QFT has the resources for providing the desired new model. And it is impossible to estimate how big a conceptual shift will be required in finding that new background theory.

The outlook on the quest for scientific understanding in an exploratory context where an unsupervised DNN powers new discoveries now comes out as follows: Being c-instrumental, we could not expect an explanation directly from the DNN. However, in pursuing steps (I)–(III) scientists would have to rely on previously established concepts, and hence be prone to assigning the wrong *meaning* to the DL

model. This would definitely hinder an extended understanding in terms of a new background theory, for:

> Research questions can be posed only with particular concepts. In the context of another conceptual scheme they may well fail to make sense, in which case they elude attention. (Steinle, 2016, p. 333)

And again, this is not mere philosophical speculation in the void, but an actual scientific problem recognized (if somewhat vaguely) by active researchers:

> Even complex problems in computer vision have been solved by *hand-crafted* features that reflect the *assumptions and expectations* that arise from common world knowledge. In geoscience and climate science, such global, general knowledge is still partly missing, and indeed, is exactly what we are seeking in research (hence, it cannot be an assumption). (Reichstein et al., 2019, p. 200; emph. added)

Note the crucial role of w-opacity here: It is exactly the ability of DNNs to 'automatically discover' important, complex features that further prediction and, at the same time, provide a basis for understanding underlying mechanisms (such as the development of secondary protein structure, or the intermediate decay of a certain massive particle). But if these features are so far not understood by humans, it is far from clear how to abstract them from a successful DNN by means of standard interpretability methods; for "the subsequent interpretation of the final state of the trained network depends on human categories expressed in natural language by the human evaluators" (López-Rubio, 2020, p. 12).

Hence, to put it in the words suggested to me by an anonymous referee, the DNN *finds* significant features, but the *translation* of these into scientific concepts is up to scientists' ability and knowledge. This is because concepts are expressed linguistically, but current state of the art DNNs do not have the ability to generate linguistic descriptions of the concepts that underlie these automatically discovered significant features.

5 Conclusion

I have argued that DNNs are c-instrumental models that harvest their success in a w-opaque way. Even though they excel as predictive tools, they thus do not deliver explanations themselves and may conceal information relevant for new-concept-formation. As I have shown, this creates the possibility of unprecedented gaps between discovery and understanding in the near future; in particular, when the following four factors are jointly present:

1. In an *exploratory experimental context*,
2. an *unsupervised* model
3. predicts an *unexpected discovery*, whose understanding
4. requires a radical *conceptual shift*.

Emphatically, I am not claiming that *only* under conditions 1.–4. will scientists face problems in gathering understanding of underlying mechanisms from DL, nor that this is *bound* to happen in case 1.–4. occur: Certainly, executing step (III) can be hard under far less drastic circumstances, and maybe the right set of geniuses, with the necessary 'exotic' ideas, are around the corner if and when 1.–4. happen (as in the quantum revolution).

However, hoping for geniuses to be around is certainly not a satisfying response to this problem, and given the current state of several big-data sciences (as well as the astonishing DL successes witnessed therein), I submit that we might plausibly face a scenario like the above in the near future.

What this means for science as whole remains to be seen: Will future scientists value prediction over explanation? Or will they develop new skills for constructing explanatory models from sparse information? There are certainly already some steps that seem to point in the latter direction: Reichstein et al. (2019), for instance, suggest a hybrid approach in which physics information is incorporated into the training, and Alvarez Melis and Jaakkola (2018) propose a framework for DNNs that are, in a sense, 'self explaining'.

However, neither of these proposals seems to make contact with the *conceptual* challenges I have raised in this paper, and I am not convinced that any approach to 'self-explanation' could deliver something like the complex, involved—and often quite *ingenious*—models that arise from advanced physical (or, more generally: scientific) theorizing.

In any case, I hope to have shown convincingly that, in a precise sense, the w-opacity and c-instrumentality of DL models indeed have the potential to profoundly 'change the face of science'.

Acknowledgements The research for this paper was funded by the German Research Foundation, as part of the research unit *The Epistemology of the Large Hadron Collider* (DFG; Grant FOR 2063). I thank audiences at the 2018 conference *The Science and Art of Simulation* at the HLRS in Stuttgart, Germany, and at the 2020 Workshop *Machine Learning: Prediction Without Explanation?* at the Karlsruhe Institute for Technology (KIT), Germany, for helpful discussions on the subject matter. I also owe thanks to Paul Grünke, Rafaela Hillerbrand, Marianne van Panhuys, Gregor Schiemann, and Christian Zeitnitz, all of whom are members of the *Poject B1: The impact of computer simulations and machine learning on the epistemic status of LHC Data*, part of the DFG/FWF-funded research unit *The Epistemology of the LHC*. I have also profited from various comments by regular visitors of the research unit's internal seminar, as well as from two anonymous referees.

Funding Open Access funding enabled and organized by Projekt DEAL. The research for this paper was funded by the German Research Foundation, as part of the research unit *The Epistemology of the Large Hadron Collider* (DFG; Grant FOR 2063).

References

Albertsson, K., Altoe, P., Anderson, D., Andrews, M., Espinosa, J. P. A., Aurisano, A., Basara, L., Bevan, A., Bhimji, W., Bonacorsi, D., Calafiura, P., Campanelli, M., Capps, L., Carminati, F., Carrazza, S., Childers, T., Coniavitis, E., Cranmer, K., David, C., … Zapata, O. (2018). Machine learning in high energy physics community white paper. *Journal of Physics: Conference Series, 1085*(2), 022008.

Alvarez Melis, D., & Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems, 31*, 7775–7784.

Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature Communications, 5*, 4308.

Basso, A., Lisciandra, C., & Marchionni, C. (2017). Hypothetical models in social science. In L. Magnani & T. Bertolotti (Eds.), *Springer handbook of model-based science* (pp. 413–433). Springer.

Beisbart, C. (2012). How can computer simulations produce new knowledge? *European Journal for Philosophy of Science, 2*(3), 395–434.

Boge, F. J. (2019a). How to infer explanations from computer simulations. *Studies in History and Philosophy of Science Part A*. https://doi.org/10.1016/j.shpsa.2019.12.003

Boge, F. J. (2019b). Why computer simulations are not inferences, and in what sense they are experiments. *European Journal for Philosophy of Science, 9*(1), 13.

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*. https://doi.org/10.1007/s11229-020-02667-3

Bokulich, A. (2011). How scientific models can explain. *Synthese, 180*(1), 33–45.

Bokulich, A. (2014). How the tiger bush got its stripes: 'How possibly' vs. 'how actually' model explanations'. *The Monist, 97*(3), 321–338.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese, 195*(12), 5339–5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass, 14*(10), e12625.

Buckner, C. (2020). Understanding adversarial examples requires a theory of artefacts for deep learning. *Nature Machine Intelligence, 2*(12), 731–736.

Buckner, C. J. (2021). Black boxes, or unflattering mirrors? Comparative bias in the science of machine behavior. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/714960

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods, 15*(4), 233–234.

Card, D., Mas, A., & Rothstein, J. (2008). Tipping and the dynamics of segregation*. *The Quarterly Journal of Economics, 123*(1), 177–218.

Chang, S., Cohen, T., & Ostdiek, B. (2018). What is the machine learning? *Physical Review D, 97*(5), 6.

Chattopadhyay, A., Hassanzadeh, P., & Pasha, S. (2020). Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports, 10*(1), 1–13.

Chirimuuta, M. (2020). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*. https://doi.org/10.1007/s11229-020-02713-0

Cho, A. (2017). Ai's early proving ground: The hunt for new particles. *Science, 357*(6346), 20.

Clark, W. A. (1991). Residential preferences and neighborhood racial segregation: A test of the schelling segregation model. *Demography, 28*(1), 1–19.

Craver, C. F. (2006). When mechanistic models explain. *Synthese, 153*(3), 355–376.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*(4), 568–589.

Davies, P. L. (2014). *Data analysis and approximate models*. CRC Press.

de Regt, H. (2017). *Understanding scientific understanding*. Oxford University Press.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science, 76*(4), 444–463.

Duncan, A., & Janssen, M. (2019). *Constructing quantum mechanics* (Vol. 1). Oxford University Press.

Durán, J. M. (2017). Varying the explanatory span: Scientific explanation for computer simulations. *International Studies in the Philosophy of Science, 31*(1), 27–45.

Durán, J. M. (2018). *Computer simulations in science and engineering*. Springer Nature.

Durán, J. M. (2020). What is a simulation model? *Minds and Machines*, *30*(3), 301–323.
European Strategy Group. (2020). *2020 update of the european strategy for particle physics*. http://europeanstrategyupdate.web.cern.ch/sites/europeanstrategyupdate.web.cern.ch/files/CERN-ESU-015-2020
Farina, M., Nakai, Y., & Shih, D. (2020). Searching for new physics with deep autoencoders. *Physical Review D, 101*(7), 075021.
Franklin, L. R. (2005). Exploratory experiments. *Philosophy of Science, 72*(5), 888–899.
Frigg, R., & Reiss, J. (2009). The philosophy of simulation: Hot new issues or same old stew? *Synthese, 169*(3), 593–613.
Giere, R. (2006). *Scientific perspectivism*. University of Chicago Press.
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. http://arxiv.org/abs/1412.6572
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.
Grimm, S. R. (2010). The goal of explanation. *Studies in History and Philosophy of Science Part A, 41*(4), 337–344.
Grüne-Yanoff, T. (2013). Appraising models nonrepresentationally. *Philosophy of Science, 80*(5), 850–861.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.
Hacking, I. (1992). The self-vindication of the laboratory sciences. In A. Pickering (Ed.), *Science as practice and culture* (pp. 29–64). The University of Chicago Press.
Harris, T. (2003). Data models and the acquisition and manipulation of data. *Philosophy of Science, 70*(5), 1508–1517.
Hasse, H., & Lenhard, J. (2017). Boon and bane: On the role of adjustable parameters in simulation models. In J. Lenhard & M. Carrier (Eds.), *Mathematics as a tool* (pp. 93–116). Springer.
Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.
Hornik, K., Stinchcombe, M., White, H., et al. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.
Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.
Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626.
Humphreys, P. (2013). Data analysis: Models or techniques? *Foundations of Science, 18*(3), 579–581.
Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters, 14*(12), 124007.
Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260.
Karaca, K. (2013). The strong and weak senses of theory-ladenness of experimentation. *Science in Context, 26*(01), 93–136.
Karaca, K. (2018). Lessons from the large hadron collider for model-based experimentation. *Synthese, 195*(12), 5431–5452.
Kasabov, N. (2019). *Time-space, spiking neural networks and brain-inspired artificial intelligence*. Springer.
Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.
Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965* (pp. 91–196). Cambridge University Press.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444.
Lenhard, J. (2006). Surprised by a nanowire: Simulation, control, and understanding. *Philosophy of Science, 73*(5), 605–616.
Lenhard, J. (2007). Computer simulation: The cooperation between experimenting and modeling. *Philosophy of Science, 74*(2), 176–194.
Lenhard, J. (2019). *Calculated surprises*. Oxford University Press.
Lenhard, J., & Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Modern Physics, 41*(3), 253–262.
Leonelli, S. (2016). *Data-centric biology*. University of Chicago Press.

Leonelli, S. (2019). What distinguishes data from models? *European Journal for Philosophy of Science, 9*(2), 22.
López-Rubio, E. (2020). Throwing light on black boxes: Emergence of visual categories from deep learning. *Synthese*. https://doi.org/10.1007/s11229-020-02700-5
Lyre, H. (2008). Does the Higgs mechanism exist? *International Studies in the Philosophy of Science, 22*(2), 119–133.
Lyre, H. (2020). The state space of artificial intelligence. *Minds and Machines*. https://doi.org/10.1007/s11023-020-09538-3
Maher, P. (1988). Prediction, accommodation, and the logic of discovery. *Philosophy of Science, 1988*(1), 273–285.
Massimi, M. (2005). *Pauli's exclusion principle: The origin and validation of a scientific principle*. Cambridge University Press.
McCrea, R. (2009). Explaining sociospatial patterns in South East Queensland, Australia. *Economy and Space, 41*(9), 2201–2214.
Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press.
Morello, V., Barr, E., Bailes, M., Flynn, C., Keane, E., & van Straten, W. (2014). Spinn: A straightforward machine learning solution to the pulsar candidate selection problem. *Monthly Notices of the Royal Astronomical Society, 443*(2), 1651–1662.
Morrison, M. (1999). Models as autonomous agents. In M. Morrison & M. S. Morgan (Eds.), *Models as mediators* (pp. 38–65). Cambridge University Press.
Morrison, M. (2015). *Reconstructing reality: Models, mathematics, and simulations*. Oxford University Press.
Morrison, M., & Morgan, M. S. (1999). Models as mediating instruments. In M. Morrison & M. S. Morgan (Eds.), *Models as mediators* (pp. 10–37). Cambridge University Press.
Napoletani, D., Panza, M., & Struppa, D. C. (2011). Agnostic science. Towards a philosophy of data analysis. *Foundations of Science, 16*(1), 1–20.
Patterson, J., & Gibson, A. (2017). *Deep learning: A practitioner's approach*. O'Reilly Media.
Poggio, T., Banburski, A., & Liao, Q. (2020). Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences, 117*(48), 30039–30045.
Potochnik, A. (2017). *Idealization and the aims of science*. University of Chicago Press.
Raghu, M., & Schmidt, E. (2020). A survey of deep learning for scientific discovery. arXiv:2003.11755 (**arXiv preprint**).
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566*(7743), 195–204.
Reutlinger, A., Hangleiter, D., & Hartmann, S. (2018). Understanding (with) toy models. *The British Journal for the Philosophy of Science, 69*(4), 1069–1099.
Royal Society and Alan Turing Institute. (2019). Discussion paper: The AI revolution in scientific research. https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
Saxe, A. M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B. D., & Cox, D. D. (2019). On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics, 2019*(12), 124020.
Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology, 1*(2), 143–186.
Schwartz-Ziv, R., & Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv:1703.00810 (**arXiv preprint**).
Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706–710.
Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034 (**arXiv preprint**).
Skansi, S. (2018). *Introduction to deep learning: From logical calculus to artificial intelligence*. Springer International Publishing.
Smeenk, C. (2006). The elusive Higgs mechanism. *Philosophy of Science, 73*(5), 487–499.

Sober, E. (2002). Instrumentalism, parsimony, and the Akaike framework. *Philosophy of Science, 69*(S3), S112–S123.
Steinle, F. (1997). Entering new fields: Exploratory uses of experimentation. *Philosophy of Science, 64*, S65–S74.
Steinle, F. (2016). *Exploratory experiments: Ampère, Faraday, and the origins of electrodynamics*. University of Pittsburgh Press.
Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.
Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A, 44*(3), 510–515.
Suárez, M., & Cartwright, N. (2008). Theories: Tools versus models. *Studies in History and Philosophy of Modern Physics, 39*(1), 62–81.
Sullivan, E. (2019). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035
Suthaharan, S. (2016). *Machine learning models and algorithms for big data classification*. Springer.
Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95–98.
Voss, H. (2013). Classification. In O. Behnke, K. Kröninger, G. Schott, & T. Schörner-Sadenius (Eds.), *Data analysis in high energy physics: A practical guide to statistical methods* (pp. 153–186). Wiley.
Wallace, D. (2020). On the plurality of quantum theories. In S. French & J. Saatsi (Eds.), *Scientific realism and the quantum* (pp. 78–102). Oxford University Press.
Wilholt, T. (2005). Explaining models: Theoretical and phenomenological models and their role for the first explanation of the hydrogen spectrum. *Foundations of Chemistry, 7*(2), 149–169.
Winsberg, E. (1999). The hierarchy of models in simulation. In L. Magnani, N. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (pp. 255–269). Kluwer Academic/Plenum Publishers.
Winsberg, E. (2010). *Science in the age of computer simulation*. University of Chicago Press.
Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation, 1*(1), 67–82.
Worrall, J. (1985). Scientific discovery and theory-confirmation. In J. C. Pitt (Ed.), *Change and progress in modern science* (pp. 301–331). Springer.