Machine Learning Discoveries and Scientific Understanding in Particle Physics: Problems and Prospects*

Florian J. Boge[†‡] & Henk W. de Regt[§]

Particle physicists have been among the early adopters of Machine Learning (ML) methods, the most notable ML systems being Deep Neural Networks (DNNs). Today, ML's use in Particle Physics (PP) ranges from the reconstruction of signals inside the detector to the simulation of events and the determination of statistical ratios in the final analysis. Most intriguingly, there is some evidence which suggests that DNNs might be able to independently acquire complex physical concepts—concepts that are relevant for the discovery and understanding of new particles and phenomena. We here argue that these two possibilities, that of discovering novel concepts per se, and that of discovering novel phenomena by means of them, pose epistemic challenges for particle physicists. In turn, we will analyse ways of mitigating these challenges, both actual and at present merely possible.

Keywords particle physics • scientific discovery • concepts • phenomena • understanding • deep learning

1 Introduction

Particle physicists face a tremendous amount of data in their experiments. Inside detectors at the Large Hadron Collider (LHC) at CERN in Geneva, numbers can ramp up to 1 billion collisions per second, leading to an incredible amount of some petabyte of collision-data per second.[1] These data are filtered down by sophisticated 'trigger systems' to just some hundreds of megabytes per second, but of course this still means a huge amount of data to be stored and analysed. Furthermore, the response of these triggers must be designed so that indeed only 'uninteresting' events get discarded. Finally, due in part to the nature of the underlying theory, the connection between theoretical predictions and recorded data is highly mediated (Morrison and Morgan, 1999), and analyses leading to discovery and measurement claims are intricate (Boge, 2021; Boge and Zeitnitz, 2021).

*Forthcoming in: Duran, J. M., and Pozzi, G. (Eds.), Philosophy of Science for Machine Learning, Synthese Library, Springer.

[†]Institute for Philosophy and Political Science, TU Dortmund University, Emil-Figge-Str. 50, room 2.247, 44227 Dortmund, Germany. florian-johannes.boge@udo.edu

[‡]Lamarr Institute for Machine Learning and Artificial Intelligence

[§]Institute for Science in Society, Chair for Philosophy of the Natural Sciences, Radboud University, Heyendaalseweg 135, 6525 AJ Nijmegen, the Netherlands. henk.deregt@ru.nl

[1]See, e.g., https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone (checked 05/23).

Given the sheer amounts of data and the complexities thus involved in data-generation, management and analysis, it is no wonder that "[p]article physicists began fiddling with artificial intelligence (AI) in the late 1980s, just as the term 'neural network' captured the public's imagination." (Cho, 2017) Over these 40 or so years, there has been a steady co-evolution of particle physics (PP) and Machine Learning (ML) methods.[2] Today, applications of ML in PP are vast and their number is ever-growing.[3] What we here, in this chapter, will focus on are a few developments which suggest that genuine discoveries in PP seem possible with ML, and that these may outstrip physicists' present-day *understanding* of the sub-atomic domain. Thus, we will here confront the possibility of discovery without (human) understanding; a phenomenon which bears close connections to the pertinent topic of scientific understanding's role in scientific progress (Dellsén, 2021; de Regt, 2017; de Regt et al., 2009; Grimm et al., 2017; Rowbottom, 2023) and to the question of whether explanation falls behind in ML (Boge & Poznic, 2021; Boge et al., 2022).[4]

In particular, we will here discuss the potential discovery of novel *concepts* by ML in PP,[5] and of novel *phenomena*. Prima facie, in either of these events, particle physicists will be at a loss regarding understanding while facing novelty in a discovery-related sense: Understanding relies on theories and models, which in turn rely on concepts. If an ML system thus conceptualizes a targeted domain in a novel way, it will be at a relative advantage compared to the scientists using the ML system (and not, or not yet, in possession of the relevant concepts). Similarly, if a phenomenon is discovered in the absence of theory, then it will not be well-understood.[6]

Indeed, it seems very much plausible that ML-based discoveries can inspire new physics-concepts (Barman et al., 2024; Iten et al., 2020; Krenn et al., 2022), and this has already somewhat happened in certain areas of physics (see Ananthaswamy, 2021). To connect these issues specifically to *particle* physics, we will use two sets of case studies, the first being concerned with the potential discovery of complex PP-concepts by Deep Neural Networks (DNNs) which have not in any way been directly given to the DNN during training (Baldi et al., 2014; Chang et al., 2018). The second will be concerned with the possibility of discovering novel phenomena by means of unsupervised learning, i.e., without specific, targeted outputs to be computed by the DNN at any given data instance (Dillon et al., 2022; Farina et al., 2020; Finke et al., 2021; Fraser et al., 2022; Heimel et al., 2019).

Assuming that understanding is one major goal of science, how will scientists remedy these sorts of situations? In this connection, we will discuss two recently proposed frameworks for 'learning from the machine' (Barman et al., 2024; Krenn et al., 2022). These proposals are aimed at mitigating the effects of novel concept-discoveries by ML (but not by humans), and hence the advantage in understanding that the machine may be said to have over human researchers. Furthermore, we will also address current limitations to discovering novel phenomena with ML, as a robust performance in this connection is actually unlikely to obtain *without* significant reliance on theory (Boge et al., MS). Thus, hybrid approaches, integrating theory and ML, are presumably the more realistic road to an ML-assisted discovery of phenomena. At the same time, these would lessen the scope and impact of 'discovery without understanding' through ML.

[2] See, e.g., https://physicsworld.com/a/ai-and-particle-physics-a-powerful-partnership/ for the reciprocal aspects of this relationship.

[3] See, e.g., the living review of ML methods in PP, https://iml-wg.github.io/HEPML-LivingReview/.

[4] In this connection, see also Chapter 7 by Páez, Chapter 8 by Buijsman, and Chapter 10 by Sullivan and Kasirzadeh in this volume.

[5] In this connection, see also Chapter 12 by Kieval and Chapter 13 by Freiesleben in this volume.

[6] Of course, this can also happen in the absence of ML, but with the vast and unwieldy data-sets of PP and like fields, the gap between discovery and understanding will certainly become significantly larger, and the challenges for overcoming it quite different.

2 ML's Discovery Potential in Particle Physics

2.1 Discovery of Novel Concepts by DNNs

The maybe most intriguing aspect of the use of advanced ML systems like DNNs in fields such as PP is that they appear to be able to acquire complex physical concepts without being instructed about them. One might be skeptical about applying such a mentalistic vocabulary to ML systems which, on a more sober view, are just complex functions on high-dimensional spaces realized on a computer and optimized over a large space of parameters.[7] Indeed, while some are optimistic that this is possible under a modest reading of 'concept' (Räz, 2023), one of us (Boge, 2024) has recently argued that it is sufficient (and indicated) to assume that DNNs can merely *emulate* concept-acquisition, and that this will even *amplify* the issues discussed below. In this chapter, however, we will act as if talk of concept-possession in DNNs was fully appropriate, and avoid deeper discussion of the thorny philosophy of mind-issues.

First note that this purported concept-acquisition is a general feature that has nothing to do with PP per se. But it acquires a special relevance therein, as we shall argue. In the philosophical literature, such an acquisition of concepts by DNNs has been most prominently acknowledged by Buckner (2018), but for our purposes, it will be most instructive to consider some actual examples from the ML and PP literature.

Let's begin with the works by Bau et al. (2017, 2018), which fall outside the scope of specifically scientific applications of ML. In a first study, Bau et al. (2017) introduced *annotation masks* for a data set of scenery-images, which contained labels for elements of the scenery on various levels (from pixel to whole image). These masks were intended to isolate conceptually meaningful patterns (say, sets of pixels representing dogs), and their accuracy was tested for with the help of verdicts by human test-subjects. Additionally, the top activated nodes in an image-classifying DNN—i.e., a DNN that is fed with an image and then spits out a label—were used to define an *activation map* over the images pixels, by blackening everything but those pixels for which a given set of nodes exceeded a very high threshold of activation. Using the matching-percentage between annotation and activation, Bau et al. (2017) could identify nodes that were reasonably specialized to concepts such as DOG—*even though the network had only been trained to classify whole sceneries.*

This is certainly impressive, as it seems as though the DNN had inevitably acquired a concept DOG as a means for classifying typically dog-containing sceneries (see also López-Rubio, 2020). A second study by Bau et al. (2018) is even more impressive in this respect. Here, Bau and colleagues investigated a *generative* DNN, which takes in white noise-images and spits out photo-realistic ones after proper training. Said generative DNN was trained to produce, e.g., sceneries involving churches, which of course also included things like trees. Walking through a similar procedure as above, Bau et al. (2018) were here able to identify, say, tree-specialized units, even though the DNN had not been specifically trained to produce trees. The novelty of this second study was that Bau et al. (2018) successively set the activations of tree-correlated nodes to zero by fiat, which led to a successive decrease of the trees in generated images. In this way, Bau et al. (2018) could show that these nodes and their activations indeed functioned *causally* in the ways that concepts supposedly do for human beings (see also Boge, 2024).

Thus, there is evidence from within the science of ML that DNNs may be said to 'possess concepts', or at least be able to mimic concept-possession by means of activation patterns correlated with data-elements that human beings would consider conceptually meaningful. However, certainly the most impressive point here is that these were concepts the DNN *had not even been 'educated' on.*

[7] See Radder (2006b, Ch. 5) as well as Chapter 13 by Freiesleben in this volume and references discussed therein for similar skepticism.

Evidence suggestive of this very same fact has been found within various physics-related applications as well. An impressive set of examples has been presented by Iten et al. (2020): A DNN with a specific architecture, called an 'autoencoder' (Figure 1 for illustration), has been shown to plausibly acquire concepts that are crucially important for understanding the underlying physics. The autoencoder has a very 'slim' intermediate layer, with only a few nodes (say two or three), so that it can store only the most salient information for the sake of predicting a particular output. When Iten et al. (2020) trained (a variant of) such an autoencoder to predict the positions of a simulated pendulum (a damped harmonic oscillator), based on past positions, they could show that the most condensed layer had specialized to the two constants determining the underlying equation (the reduced spring and damping constants), *without ever having 'seen' said equation.* Similarly, when said DNN was trained to predict the positions of planets in our solar system and of the sun from a geocentric perspective, they found that nodes in the condensed layer had specialised to angles describing the trajectories in a *heliocentric* view.

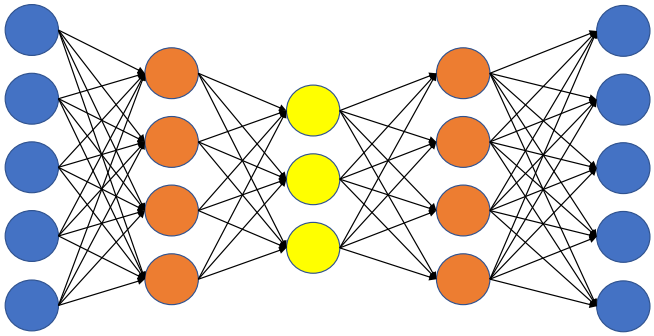

Figure 1: Generic depiction of the 'autoencoder' architecture. The yellow, slim layer is called the 'bottleneck'; the layers preceeding it an 'encoder', those succeeding it a 'decoder'.

These futuristic-sounding observations have a correlate also within *particle* physics. In order to expound on this, let us briefly recapture the very basic ideas surrounding particle physics events, such as those going on at the LHC in Geneva. In order to be able to detect the particles they are interested in, particle physicists smash together other particles that they know how to isolate and exert control over. At the LHC these are protons, accelerated to peak collision energies of almost 14 TeV. The interaction between these is conceived of in terms of interactions between the quarks and gluons, collectively referred to as 'partons', that the protons are 'made of'.

Typically, multiple partonic interactions will take place at once, and there may even be several proton-proton collisions taking place in close proximity at roughly the same time. This leads to additional activity, sometimes collectively called an 'underlying event'. Strictly speaking, all these 'events' are of a quantum-mechanical nature, which means that there can be interference between several distinct processes of partonic interaction (Passon, 2019; Schwartz, 2021). However, as a matter of fact, there are theoretical reasons to think that, at very high energies, such interactions can be treated in close analogy to statistical ensembles of 'classical' interactions between tiny particles (Schwartz, 2014). Another aspect of the quantum field-theoretical nature of the underlying theory is the fact that such events correspond to the *annihilation* of the interacting particles and the *creation* (and subsequent decay) of particles of interest. Many of the quarks contributing to the overall activity are spontaneously created from the vacuum. Furthermore, as the particles so produced lose energy, they are expected to form 'hadrons', i.e., multi-quark particles (or bound-states between several quarks), through complicated and only partly understood processes that also involve the spontaneous creation of quark-antiquark pairs from the vacuum. It is these hadrons, or even the particles they quickly decay into, that are

then actually measured in the detector (Boge and Zeitnitz, 2021).

To provide clear evidence for the intermediate existence of a given particle, or even to measure some of its properties, physicists need to match the load of data produced in multiple events to theoretical models and their predictions. Typically, they do not use the data directly 'given' to them by the detector for this. As mentioned above, electric currents received from the detector will first be filtered down by means of a sophisticated 'trigger system' (also Karaca, 2018), intended to store only events that are potentially of interest, so that there even is a manageable amount of data left. The data that survive this procedure are stored as 'event records', specifying the activity in the detector in terms of lists of numbers (Albertsson et al., 2018; Delfino, 2020). They are also referred to as the 'raw data', and these will be processed further into 'low level' data specifying the momenta, scattering angles, and some further identifying properties of the particles actually measured in the detector.

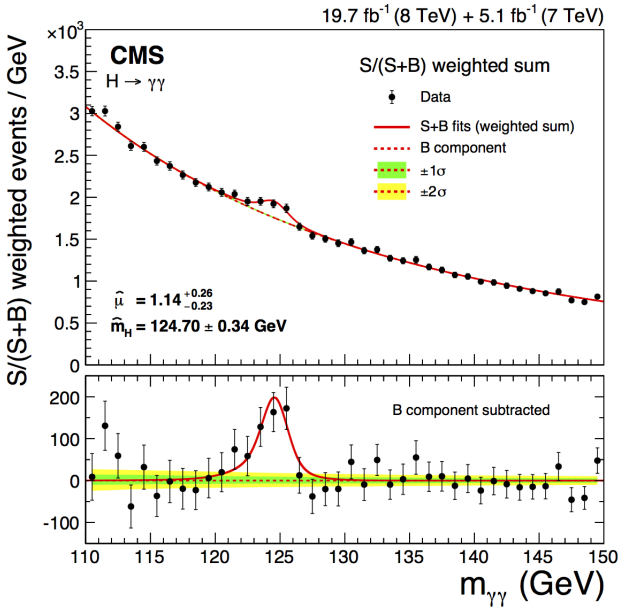

Figure 2: Statistical displays of the Higgs-observation in 2012. Taken from (CMS, 2014) under a CC BY 4.0 license. Colour available online.

However physicists usually even go one step further and reconstruct, say, the (invariant) masses of intermediate particles from these 'low level data', by appeal to physical laws such as the relativistic energy-momentum relation. Data so derived, through these and further intermediate steps, are then referred to as 'high level data' (Delfino, 2020). The key reason for using high level data is that this representation is often most discriminating between a null hypothesis, typically called the 'background only' hypothesis in the jargon, and the alternative (that there is something unexpected there). The classic image is that of a 'bump', as displayed in Fig. 2, indicating the intermediate production of higgs bosons, decayed into two photons, in addition to a 'background' of pairs of photons to be expected also in the absence of higgs bosons.

Remarkably, some DNNs trained on low level data to classify events as either 'signal' or 'background', and to so to help even determine whether there is an excess of 'signal' data, did not profit from being handed also the high level information (Baldi et al., 2014). Even more impressively, Chang et al. (2018) found that, when the data were prepared so that the information on higher level variables such as the invariant mass was removed, an initially very successful DNN started to fail miserably. In just a little more detail, consider the histogram displayed in Fig. 3. In the two left most panels, one can see two histograms displaying the frequency of

events with a specific reconstructed invariant mass. As is readily seen, the top histogram here features the characteristic bump indicative of a new particle. However, manipulating the data, by weighting each column in this histogram inversely by its height, one obtains the lower of the two histograms.[8] Obviously, no 'bump-information' for the mass distribution is contained in the data anymore. This procedure was called 'data-planing' by Chang et al. (2018), as the result is a flat, uniform distribution.

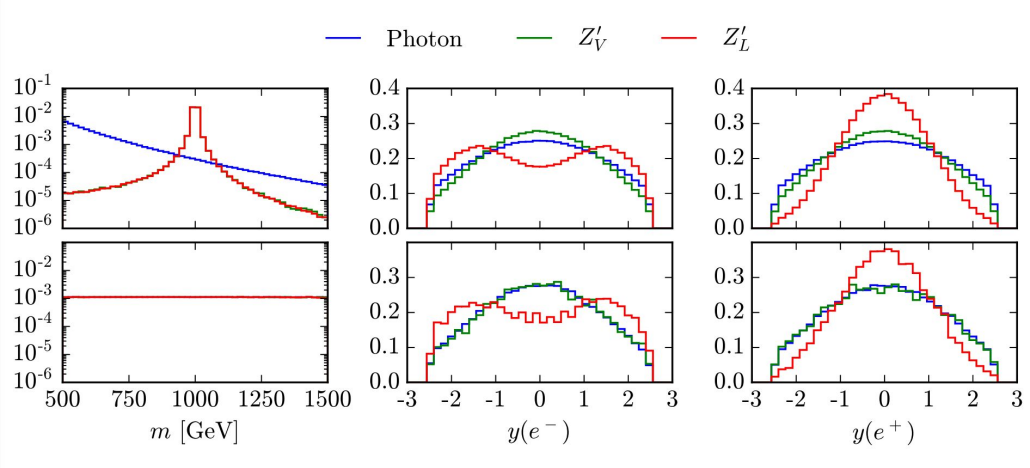

Figure 3: Illustration of the data-planing procedure used by Chang et al. (2018) to rob a DNN of the information on higher level variables contained in the data. Blue data correspond to the 'background only' simulation, red and green data to two distinct versions of a hypothetical new Z'-particle. As can be seen, the effect of planing away the mass information only leads to swift changes in the other plotted quantities ($y(e^{\pm})$) for the Z'_L-case. Reproduced from Chang et al. (2018) under a CC BY 4.0 license. Colour available online.

As a result of this planing-procedure the DNN lost its ability to discriminate signal from background. However, the crucial point is only seen by considering the remaining histograms (middle and right), where the top one always corresponds to the original histogram, *before* the planing, and the bottom one to the result of planing. These are histograms for *different* quantities and, as is readily seen, the changes in these histograms are rather swift. Hence, it is unclear whether one would even *notice* a specific loss of information *unless one was already in possession of the relevant concept* (invariant mass). In other words: This study suggested quite vividly that the relevant DNN had recovered the mass information on its own, without having been instructed about masses at all.

These fascinating observations hint at the possibility that DNNs and other complex ML systems may acquire concepts on their own. However, combined with another observation, this even yields the possibility that DNNs may develop (or maybe 'emulate') concepts that we, as human researchers, are not even in possession of.

In this connection, consider the prominent case displayed in Fig. 4, in which a DNN is made to fail by adding some dedicated noise to an originally well-classified image of a panda. After the addition of said noise, the DNN recognises the image as displaying a gibbon, and even more confidently than was the case with the panda before. Images (or more generally: data) of this sort are usually called *adversarials*. Next to data consciously created to fool the DNN, as in the panda case, a more permissive reading of the term comprises any sort of image (or datapoint) that is easily recognizable for a human but surprisingly misidentified by a given DNN.

[8] Actually, $10^{-x} \times 10^{x} = 1$, so the final histogram also includes a normalization-step.

Figure 4: Famous example of an adversarial, taken from (Goodfellow et al., 2014). Color available online.

Such a situation may be brought about, say, by including an unusual background to an object (Hendrycks et al., 2021), or by displaying it in a particular pose (Alcorn et al., 2019).

Crucial for us here is the reason *why* DNNs fail on such images or data. As Szegedy et al. (2013) noticed early on, the presence of adversarials seems to be fairly general across DNN architectures and data-sets used for training. This suggests that there might be generic reasons related to the very functioning of DNNs that lead to the possibility of finding or generating adversarials. Indeed, recent work by Zhang et al. (2019) suggests that adversarials are closely connected to what is atypical *for* the DNN, relative to its training, in the sense of being 'distant' from the training set in a specific metric. This metric was defined by comparing what a DNN encoded in its hidden layers when confronted with a new example to what it so encoded on the training data. Thus, if $h(x_{\text{test}})$ is the activation of some specific hidden layer of a DNN on a test-image, x_{test}, and $h(x_{\text{train}})$ is the corresponding activation on some training image, then the distance in question was defined as the average distance between $h(x_{\text{test}})$ and the k nearest $h(x_{\text{train}})$ for x_{train} from the training set in an ℓ_p metric. Images 'far from' the training data in this sense were, however, usually not in any specific sense atypical for humans. When such DNN-atypical images were just slightly distorted, this resulted in an adversarial, while still looking perfectly good to humans.

This last point is key here: Computer scientists refer to the functions DNNs compute within their hidden nodes as 'features'; the suggestion being that the given number computed corresponds to some property the DNN has managed to abstract from the data (also Buchholz, 2023; Buckner, 2018). And it thus seems that the features recognized by DNNs may differ markedly from those recognized by humans: An image that is atypical in terms of the features recognized by the DNN is easily turned into something it misidentifies, but neither the atypicality nor the strong response to these swift changes are correlated with corresponding human responses.

Furthermore, a study by Ilyas et al. (2019) recently suggested that, while these features so abstracted by DNNs may not be very robust in the sense that a small amount of noise can often destroy them, they may still be well-generalizing across various data-sets. This was shown by Ilyas et al. (2019) in the following way: Defining the *utility* of a certain feature by the degree to which it is positively correlated with a desired output in classification, and its *robustness* by the degree to which it remains useful under perturbations, Ilyas et al. (2019) constructed a data-set in which only *non-robust*, but still *useful* features remained. Then, training a DNN on such a data set and testing its performance on a standard test set (with robust features included), the DNN could be shown to exhibit a very good performance. So it seems that the DNN recognizes, and exploits, the very same (non-robust) features in the test-set.

Assuming that it is, hence, meaningful to talk as if DNNs develop concepts for the features they 'recognize', a fascinating possibility (recently discussed by Buckner, 2020, in only slightly different terms) arises: DNNs could develop concepts that, while being fairly incomprehensible for humans, are nevertheless scientifically productive.

To understand this suggestion a little better, consider Goodman's (1955) bleen and grue. To us, being 'green and observed before t or else not so observed and blue' does not seem like a meaningful property. But remember that green and blue can be similarly defined by means of bleen and grue. Hence, the switch between the two sets of concepts has much in common with a change of coordinate system, or frame of reference. It thus requires additional justification to argue for the objective preferredness of one frame of reference over the other.

Now, as Buckner (2020, 3) points out,

> Quine[9] famously suggested that [...] evolution has shaped our perceptual and cognitive faculties [in such a way that] certain features [...] jump out at us as natural candidates for investigation [...].

Combined with Ilyas et al.'s observation that the 'non-robust' features perturbed in the generation of adversarials are nevertheless well-generalizing, this suggest that DNNs and like ML systems rely on features, or even entertain concepts, that we would consider 'non-natural'—somewhat like bleen and grue. However:

> If scientific investigation would become more productive by tracking non-natural features [...] then even Quine would be likely to embrace this alternative route to scientific progress. (Buckner, 2020, 3)

In other words: It seems very much possible that DNNs and similar ML systems might have an advantage on us in conceptualizing vast and unwieldy data sets, not just despite the fact that they seem to conceptualize very differently, but even *because* they do so: Maybe the 'features' with the greatest predictive impact are very hard for humans to even grasp as features (i.e., as properties of the data or of the underlying objects), but can easily be utilized by a DNN or other ML system. However, assuming that such a situation actually occurs – which could in principle very well happen in the exploratory PP-research to come in the near future (Boge, 2022) – where would that leave us as human researchers in the quest for scientific understanding?

2.2 Discovery of Novel Phenomena by DNNs

The discovery of novel concepts by ML is certainly one interesting path to discovery that might well be around the corner in PP, but there are also other (not entirely unrelated) paths. Duede (2023) has recently investigated the discovery of novel theories with the help of DNNs, which seems at least partly connected to the discovery of concepts discussed above: If the DNN discovers relevant concepts, we might treat it as a source of inspiration for our own theorising (more on that in Sect. 3.4). However, we here want to turn to the discovery of novel *phenomena*, in the *absence* of theory. As a matter of fact, this issue is partly *connected* to the above discussion on the possibility of novel concepts, but in rather subtle ways.

To unfold these issues, consider the autoencoder DNN depicted in Figure 1 again. Using such an autoencoder, particle physicists have devised a basic method for finding anomalies without building (strongly) on theory. However, to understand this suggestion, we first need to consider what an 'anomaly' actually is. First, recall the incredible amounts of data in PP and the highly mediated connection between these and the theory. Could a single, unexpected data-point tip us off as an 'anomaly' in PP? Certainly not. In PP at least, but presumably in science more generally, a recognizable anomaly will have to be of a statistical nature, reflecting an excess of data over the theory-guided expectation. Anomalies in PP are thus scientific *phenomena*: "something public, regular, possibly law-like, but perhaps exceptional." (Hacking, 1983, 222)

[9] See (Quine, 1969).

Furthermore, the word 'anomaly' suggests a deviation from a law of nature; so anomalies should be recognized as something that is inconsistent with accepted theory (containing the candidate laws for a given domain of inquiry). Hence, it seems that the very idea of recognizing anomalies in a theory-independent way is flawed: What counts as an anomaly is *defined* by means of existing theory. This is certainly correct in principle, but it need not imply any particularly *interesting* sense of theory-dependence: If the DNN is trained on data that are assumed to be well-understood but then indicates something odd on new data, this means that the oddity only relies on the acknowledged theory, not on any rivals.

Such a methodology is certainly at odds with standard methodology in PP, which is strongly theory-driven. Recall how PP's Standard Model (SM) comes with certain limitations: It does not feature a model of gravity or, relatedly, dark matter (Martens and Lehmkuhl, 2020), and has many suspicious fine-tuning properties (Rosaler and Harlander, 2019). Hence, a major task in present-day PP is to find signs of physics 'beyond the SM' (BSM). Yet, "the search for many of the favoured BSM scenarios has been unsuccessful" (Butterworth et al., 2017, 2), and so new strategies are indicated. Strategies that particle physicists have explored comprise probing the room left for favoured BSM scenarios in the space spanned by their parameters by measurement uncertainties around the SM predictions (Butterworth et al., 2017), using simplified models that include certain BSM properties (McCoy and Massimi, 2018), and extending the SM by exploring rather arbitrary new combinations of its operators (Bechtle et al., 2022). However, these approaches clearly presuppose not only the SM but also some candidate, or preliminary, *rivals*. In contrast, using a DNN to probe the data promises a much greater extent of theory-freedom.

So far, this is all certainly consistent with the traditional ways in which anomalies have promoted progress: Remember how Kuhn (1970, 52–3) held them responsible for paradigm-shifts:

> Discovery commences with the awareness of anomaly [and] continues [...] until the scientist has learned to see nature in a different way [...].

Clearly, one need not be a Kuhnian to acknowledge that anomalies matter to progress though: More moderate thinkers like Lakatos (1970, 1976) also reserved an important role in theory change for them, and one of us (de Regt, 2020) has recently shown how the resolution of an anomaly in the kinetic theory of gases led to an increase in understanding – and thus to scientific progress.

However, as was explained above, even recognizing an anomaly becomes very difficult when the data are as vast, and the theory is as strongly mediated by models, as is the case in PP. Thus, the crucial point is that theory may even be insufficient to say when something is *not* consistent with accepted laws.

So DNNs promise as theory-independent a method for recognizing anomalies as possible, and in domains where this would otherwise hardly be humanly feasible. Let us see in more detail how it works. In two benchmark studies, Farina et al. (2020) and Heimel et al. (2019) could demonstrate some astonishing successes with autoencoders on simulated data. Focusing on the study by Farina et al. (2020), the task faced by the autoencoder was to recognize anomalous 'jet images', which were chosen to stem from top-quark events.

To unpack the physics a little here, recall how quarks inevitably form hadrons (multi-quark particles like protons and neutrons) at lower energies – a phenomenon known as 'quark confinement'. The related process is complicated, and involves particles spontaneously created from the vacuum, as well as decays of semi-stable particles that form over short times (Boge and Zeitnitz, 2021). However, what one measures in the detector are actually the products of this 'hadronization process'. The distribution of tracks of such particles in the detector from events involving the annihilation of (almost) free quarks is identifiable by its characteristic, conical shape. Such a cone of tracks is referred to as a 'jet'. Furthermore, jets from distinct types of quark-events may

exhibit different sub-structures (i.e., different internal groupings of tracks jointly forming a jet into sub-jets). The identification of both jets and their sub-structures is non-trivial. However, jet-substructures can be made visible by 'looking into the jet from above', i.e., displaying an image of the distribution of energy-momentum inside the detector when looked onto from above the center of the jet, down into the jet's core. For humans to recognize characteristic structures in this way, it is necessary to average over some tens or even hundreds of thousands of such images. Obviously, this requires first being able to *sort* them according to the particle-type they stem from.

Clearly, such an approach doesn't make much sense if one is searching for an anomaly. Hence, the approach pursued with autoencoders is slightly different. Instead of averaging over images, the autoencoder is trained to reconstruct individual jet-images which can be assumed to be consistent with the known physics. If after this training, the autoencoder is unable to reconstruct a given jet-image well, this is an indication that it contains anomalous features and thus may be indicative of an as yet unknown ('anomalous') physics process.

Surprisingly, in benchmark studies with simulated data wherein top-quark jets were used as the anomaly-model and the remaining jets from quantum chromodynamics as the 'background' (i.e., the data consistent with known physics), an autoencoder trained on a data set that was contaminated with up to 10% of anomalous jet-images could be successfully used for detecting anomalies after training (Farina et al., 2020). Hence, on the face of it, this seems like a very promising approach to finding anomalies in as theory-free ways as possible, and in domains where one could otherwise hardly find them.

There is an important catch with the theory-independence of this method, but before we discuss it, first consider the connection to concepts. Recall that the concepts discovered by DNNs (if any) may actually not be closely aligned with human concepts, and that this could actually be the factor that gives them an edge in scientific discovery (Buckner, 2020). As a matter of fact, on the face of it this is actually what happens in unsupervised anomaly detection (see Boge et al., MS): The autoencoder reacts to subtle changes in the image that are barely even recognizable to humans. Because in the condensed, bottleneck-layer, it can only focus on a small number of salient features which it abstracts from the data – but which typically won't correspond to anything meaningful for humans –, a failure to reconstruct anomalous images well indicates a difference in features *recognized by the DNN*. As we will explain in more detail in Sect. 3.2 the recognition of certain features (or even objects) defining a new phenomenon requires the presence of appropriate concepts, which may then give rise to new theories. But if this is correct, research into the discovery of phenomena by DNNs and into the concepts developed by them might ultimately co-align.

Here comes the catch though: It is far from clear that this method for recognising anomalies in such a theory-free way can be made robust (Boge et al., MS). The evidence for this comes, again, from PP. In particular, in a study by Finke et al. (2021), the autoencoder was first trained in the same basic way as in the study by Farina et al. (2020), though without contamination by top-quark anomalies. However, Finke et al. (2021) then turned things around and trained their autoencoder to reconstruct top-images (making them the 'background') and to recognize the remaining quantum chromodynamics-jets as anomalies. Surprisingly, in this latter set-up, the autoencoder performed "worse than picking anomalies randomly." (Finke et al., 2021, 3)

Indeed, it turned out that the excellent performance in the original setup was probably due to the fact that top-jets are simply *more complex*, in the sense that they contain more pixels that are important to their structure, and that they simply have more structure in the first place (ibid.). Hence, the seemingly successful autoencoders in the studies by Farina et al. (2020) and Heimel et al. (2019) may have actually *only learned to reconstruct simple images well*, while remaining unable to reconstruct complex (and *coincidentally*: anomalous) ones.

Again, the most crucial point for us here must be extracted with some additional effort. To this

end, first note that there was a certain dependency of the observed effect on the training regimen. In most of these studies, the 'loss' encountered by the DNN at any given image was quantified by the average number of pixels it failed to reconstruct correctly. However, it is not difficult to see that this can be related to bright pixels becoming more important: The brightness of certain pixels corresponds to them having 'high values', and so reconstructing these can quickly raise the average. However, given the capacity constraints imposed by the bottleneck, it will be difficult for the autoencoder to reconstruct images with 'too many' bright (and thus: salient) pixels.

Second, Finke et al. (2021) tried out several ways of mitigating the problem that were in no obvious way theory-dependent. For example, they tried to 'smear out' certain pixels, so that the autoencoder would be less focused on individual, salient pixels prominent in the simpler images. However, such strategies turned out to be quite unsuccessful (ibid., 18).

Third, there were additional studies done in which similar observations could be made, but with slightly different upshots. In particular, Fraser et al. (2022) tried out different combinations of loss-function (determining the training) and anomaly-measure (determining how badly a DNN would be able to reconstruct an image *after* training, i.e., indicating how anomalous it was). As it turned out, though, different combinations were here favourable for different types of (simulated) background and anomalous data. Hence Fraser et al. (2022) concluded that "without a signal model in mind, optimizing analysis strategies is hard to do in a principled manner" (13), or even that "one cannot optimize without a signal model in mind" (8).

One might suspect that all this is due to the architectural constraints imposed by the autoencoder and that a different architecture – without a bottleneck-layer – might fare better as an anomaly detector. However, note that the very idea of anomaly detection with ML is predicated on the presence of a bottleneck: Any sufficiently rich DNN would be trivially able to reconstruct any image, as it could learn the identity mapping. Hence, the bottleneck is *necessary* for at least the traditional ways of performing anomaly detection with ML. Furthermore, a larger bottleneck might be necessary to reconstruct complex anomalies, but figuring this out would require *preconceptions* of the type of anomaly in question on the side of the researcher – exactly our point.[10]

There appears, in other words, to be a certain trade-off between the robustness of the method and its theory-independence (Boge et al., MS): Methods which start from very innocent ideas that require little input from theory will turn out to be sensitive to many changes in, say, the types of data they encounter, the ratios between different such types, and the correlation between choices of loss-function for the training and the type of data they will be confronted with. They can hence not be said to perform robustly. As it turns out, this situation can be remedied (or at least mitigated) by appeal to expectations of what potential anomalies might look like – thus re-introducing a fairly strong theory-dependence.

In sum, it seems that the possibility of discovering of novel phenomena with DNNs must, at least at present, be taken with a grain of salt. Whether scientists come up with methods that are not subject to the same robustness-theory trade-off only time can tell (however, see also Boge et al., MS, for a principled argument exposing the obstacles). We will return to this issue below.

3 Understanding (and) ML's Impact on Particle Physics

3.1 Scientific understanding: a philosophical account

Above, we have indicated tentative problems for gaining scientific understanding in PP when discoveries – of either concepts or phenomena – are partly ML-driven therein. However, what

[10] There are several other, less popular methods (e.g. Mattia et al., 2021), but at the level of loss functions and training data, it remains unclear whether they can do without significant input from theory.

do we mean by 'understanding' here, and what are the connections to phenomena and concepts, respectively?

The nature of scientific understanding has been a central topic of debate in philosophy of science since the turn of the millennium. In contrast to traditional philosophical work on scientific explanation (a notion that is obviously related to understanding), the increasing philosophical attention for understanding is motivated by the practice-turn in the philosophy of science. Study of scientific practice (and its history) clearly shows that understanding plays a key role in the process of science, contrary to what traditional philosophers of explanation believed. An influential example of such a practice-oriented analysis is De Regt's (2017) contextual theory of scientific understanding, which is based on historical case studies of scientific development, especially physics, and on recent insights that philosophers of science have derived from studying scientific practice in general. In our chapter we will use this theory as a starting-point for our analysis of the role of ML in discovery and understanding in PP.

De Regt's core idea is the thesis that scientists achieve understanding of a phenomenon P if they construct an appropriate model of P based on a theory T. If it is empirically adequate, the resulting model constitutes an explanation that provides understanding of P. While this may not yet sound very different from traditional accounts, De Regt's claim is that study of the history and practice of science teaches us that only if theory T is intelligible to scientists is it in fact possible to construct a successful model-based explanation. Hence, the following Criterion for Understanding Phenomena (CUP; de Regt, 2017, p. 92):

CUP A phenomenon P is understood scientifically if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency.

The key term in this criterion is 'intelligible', but what exactly is intelligibility? De Regt (2017, p. 40) defines it as follows:

Intelligibility: the value that scientists attribute to the cluster of qualities of a theory T (in one or more of its representations) that facilitate the use of T.

Note that intelligibility is not an intrinsic property of theories, but a context-dependent value: whether a theory is intelligible to scientists depends on contextual factors such as their skills and background knowledge. Why do scientific theories need to be intelligible to the scientists who use them? Again, this follows from study of scientific practice, especially from the work of philosophers like Cartwright (1983) and Morrison and Morgan (1999), who highlighted the crucial role of modelling in explanatory practices. On their model-based account of scientific explanation scientists acquire understanding of the phenomena by constructing models, which 'mediate' between relevant theories and the phenomenon-to-be-explained. Constructing such mediating models involves pragmatic judgments and decisions, since models do not follow straightforwardly from theories and neither from the empirical data. In particular, suitable idealizations and approximations need to be made, which cannot be deduced from theory or data. De Regt argues that the construction of such models – which provide explanatory understanding of phenomena – requires theories that are intelligible in the sense defined above. Only if scientists' skills to work with the theory allow them to make suitable pragmatic judgments, will they succeed in constructing explanatory models. In short, understanding a phenomenon on the basis of T depends on an appropriate combination of skills of the scientist S and qualities of T.

A test for intelligibility can described by the following criterion (de Regt, 2017, p. 102):

CIT$_1$ A scientific theory T (in one or more of its representations) is intelligible for scientists (in context C) if they can recognize qualitatively characteristic consequences of T without performing exact calculations.

This is a sufficient but not a necessary condition for intelligibility. This is why there is a subscript 1 in CIT_1, since there might be other criteria for intelligibility. CIT_1 holds primarily for theories with a mathematical formulation and is thereby particularly suitable for physics. The key aspect of this condition is the ability to derive (qualitative) consequences.

Although De Regt's contextual theory is not the only one on the market, we believe it is a good starting-point for analyzing the prospects and limitations for using ML methods to achieve understanding in particle physics. First, it is an influential account that is widely accepted as capturing basic intuitions regarding scientific understanding, and second, it has been geared towards the practice of physics. It is immediately obvious that the contextual theory applies specifically to human scientific understanding: it has been developed via analysis of how human scientists achieve understanding in practice, and its emphasis on the role of skills and context highlights the characteristically human nature of understanding. This is captured by its central notion of intelligibility (of theories or models), which relates human scientists (with particular skills), in particular historical, disciplinary and social contexts.[11]

Given the fundamentally human nature of scientific understanding the question arises whether ML-driven science can conform to extant criteria for understanding, especially the intelligibility criterion. This is prima facie not the case. To begin with, according to the contextual theory, (human) scientific understanding requires theories as a basis for constructing models of the phenomena to be explained. However, both theory and model are representational devices that are allegedly completely absent in ML systems. To be sure, such systems (e.g. DNNs) are called models as well, but they are models in a specific sense that does not conform to the idea of models as mediators between theory and phenomena; instead they are complex mathematical representations that learn from input data in order to make predictions, but remain black boxes that are opaque and thereby unintelligible to humans (Boge, 2022; Facchini and Termine, 2021). Hence, DNNs seem to be incapable of generating scientific understanding in the sense of CUP, because their predictions are not based on scientific theories (indeed, they are data-driven rather than theory-driven) and they do not provide intelligible models of the phenomena.

This conclusion is not merely a peculiar feature of De Regt's theory: it resonates with the views of many philosophers and scientists. The fact that DNNs are unintelligible black boxes has generally been regarded as an obstacle to their capacity for providing understanding. Thus, Chirimuuta (2021, 787) argues that in computational neuroscience the use of artificial neural networks for modeling neuroscientific phenomena entails "a trade-off between predictive accuracy and the ability of models to confer understanding". She suggests that intelligibility – which she rightly claims to be a "human-relative virtue" – should not be given up in favor of mere prediction and control. In a similar vein, Greif (2022, 130) evaluates DNNs with regard to 'model intelligibility' and concludes that "DNNs, like ML approaches more generally, are neither designed for [scientific explanation and understanding], nor can they be recruited for it in a similar way to analogue or more traditional computer models. Their ability to master complex cognitive tasks in ways that are in part beyond human comprehension actually testifies to that".[12]

To be sure, some have argued that one can acquire scientific understanding of phenomena without model intelligibility. Kästner and Crook (MS, ch. 19, this volume), for example, claim that the arguments above rest on a misguided conflation of understanding models and 'model-induced understanding' of phenomena. They adopt a conception of scientific understanding that

[11] It is not a coincidence that philosophical interest in the notion of scientific understanding emerged only after the traditional objectivist approaches in the philosophy of science (and in the debate about scientific explanation in particular) had been abandoned and the subjective and pragmatic nature of understanding was not regarded as detracting from its epistemic importance.

[12] Both Chirimuuta and Greif adopt de Regt's notion of intelligibility. Ultimately, we think that even considering ML models as 'models' in a very literal sense is a non-starter in fields outside computational neuroscience. Hence, some of the positions discussed here may actually rest on a confusion. We will make all this clearer in Sect. 3.3.

incorporates elements of De Regt's contextual theory, namely that understanding consists in a qualitative grasp of the target phenomenon P, involving the abilities to make coarse-grained predictions and to control and intervene on P. However, they apply this directly to the phenomena rather than to representations thereof by means of theories and models, arguing that understanding a phenomenon cannot be reduced to understanding a model, since "real-world scientific understanding of complex phenomena frequently depends upon facts and details extrinsic to any individual model". Subsequently, Crook and Kästner compare simple models with complex ones and claim that "partial understanding of a complex but accurate model can lead to greater understanding of a phenomenon than complete understanding of a simple but inaccurate model". However, even if we grant this much, we submit that in the limit where complex models become completely unintelligible black boxes (as appears to be the case with DNNs), understanding of the phenomena has vanished for human scientists because they have lost their qualitative grasp in terms of abilities to predict, intervene and control on the phenomenon.

Crook & Kästner are surely right that model intelligibility and model-induced understanding differ and should not be conflated. However, interaction between the two is essential: it constitutes a feedback mechanism between constructing and using explanations that produces new scientific knowledge and thereby advances scientific understanding. This mechanism shows that understanding, as de Regt (2017, 44-47) states, is both "a means and an end". Rather than a conflation, it contains a 'virtuous circularity': scientists employ their initial understanding of theories (read: model intelligibility) to construct explanations that, if empirically successful, produce understanding of phenomena. Next, they may apply, extend and refine their knowledge of the target systems using the same skills that were needed to construct the original explanations (the skills associated with model intelligibility).

We conclude that scientific understanding requires theories on the basis of which scientists construct (model-based) explanations of target phenomena that generate understanding. At first sight it seems that while DNNs may excel in predictive power, their understanding-providing power is low because they lack intelligible theories. In the next subsection we will examine this tentative conclusion more closely by focusing on the role of concepts and the nature of phenomena.

3.2 Concepts, Phenomena, and Scientific Understanding

Above, we have argued that (intelligible) theories are required for achieving scientific understanding, but we have not yet given a characterization of what a theory is. De Regt (2017, 30-32) adopts Giere's (2006) notion of theory as a collection of principles that provide the basis for the construction of models of parts/aspects of the real world. This is a rather minimal characterization that does not put strong constraints on the notion of theory. But it does entail that the fundamental building blocks of theories are *concepts*. At this point there seems to be an opening for DNNs to develop theories after all, since – as we have seen in Section 2.1 – DNNs abstract 'features' from data, a process that can be regarded as the acquisition of corresponding concepts. However, these concepts can differ wildly from ordinary (human) concepts and may therefore be unintelligible to humans. Hence, even if DNNs develop theories on the basis of their 'concepts' and acquire understanding via these theories, this understanding may not be accessible to humans. In order to explore the ways in which DNNs' concepts (and hence theories) agree and/or differ from that of human concepts and theories, let us first consider human scientists' concepts and their relation to theories, phenomena and understanding. In Section 3.4 we will then examine whether, and if so how, conceptualizations by DNNs can be made intelligible to humans.[13]

Concepts are a thorny subject in the philosophy of mind and cognitive science. While there appears to be vast agreement that concepts are central to philosophy and human cognition

[13]For related discussions, see also the Chapter 12 by Kieval and Chapter 13 by Freiesleben in this volume.

in general, there is little agreement on the details: a variety of accounts exists of the nature and structure of concepts, of concept acquisition, and of the dynamics of conceptual change. We cannot possibly cover this debate here and will hence restrict ourselves to discussing a few insights that are relevant to our project.

To begin with, there are at least three views regarding the nature of concepts, regarding them as either (1) mental representations, (2) abilities (peculiar to cognitive agents), or (3) abstract objects (see Margolis and Laurence, 2023). For our purposes, (1) and (2) are most relevant. If concepts are assumed to be *mental* representations, DNNs by definition cannot possess them. However, DNNs can mimic concept possession by extracting features from data sets, which can be regarded as representing target phenomena. Structuring observation, detecting empirical regularities, and classifying objects are representational roles that DNNs may fulfill. The roles of enabling inferences and explanations better fits the ability view of concepts.

As regards the structure of concepts the classical view is that concepts involve necessary and sufficient conditions. This strict definitional view is challenged by the looser 'prototype theory', inspired by Wittgenstein's idea of 'family resemblance' (see Hampton, 2006). Prototype theory is congenial to the ability view of concepts insofar as it seems to reduce concept-possession to "having a bundle of inferential capacities" (Fodor, 1994, 108). However, a general worry raised by Fodor (1994, 109) – one that we share – is that the *identification* of concepts with (inferential) abilities goes too far, as this would likely defy the compositional structure of concepts.

Prototypes, furthermore, are not to be misconstrued as 'typical examples', but are "the centers of clusters of similar objects", which "allows for the representation of different possible values of relevant features such as that apples can be red, green, brown, or yellow [...]." (Hampton, 2006, 80) It is hence difficult to pin them down as mental representations.

In response to the classical and the prototype theory, a third view has been advanced that solves problems of its predecessors and is therefore gaining popularity in cognitive science: the 'theory theory' of concepts (Carey, 2011). On this view, concepts should not be considered in isolation but as part of more general theories that relate various concepts: by being part of a coherent theory concepts acquire their meaning and function, both in science and in everyday life. Hence, the theory theory is fully in line with our claim above that scientific theories are networks of concepts.

As mentioned, we cannot settle these debates here. For our purposes three conclusions are relevant. First of all, concepts are crucial for, and perhaps even inherently intertwined with, *theories*. Second, concepts can fulfill *representational* roles. Finally, concepts appear to *provide* (though they are not to be identified with) certain *abilities*, such as "recognition, naming, inference, and language understanding." (Piccinini, 2011, 179) Turning to the role of concepts in science, important abilities that concepts may provide are: enabling the recognition of phenomena, as well as enabling inferences and explanations, and thereby – as we will argue below – achieving *understanding* of those phenomena. These conclusions align with Arabatzis (2019, 86), who lists a variety of roles that concepts can play in scientific practice. For our purposes, the following roles are relevant: they can structure scientific observation; enable the detection of regularities and empirical laws; go hand in hand with the classification of objects; enable inferences about the objects they refer to; and enable the explanation of phenomena via hidden entities and mechanisms.

That concepts are crucially involved in the recognition of phenomena has been acknowledged since logical empiricism was challenged by philosophers such as N.R. Hanson and Thomas Kuhn. They famously argued that (scientific) observation is always 'theory-laden', because it is in part determined by the conceptual framework of the observer. In this vein, Koningsveld (1973, 5, emph. altered) writes: "*what* we observe, the empirical datum, is co-constituted *as that datum* by the theory or the concepts through which we observe." In these classic debates no distinction was made between observations, data and phenomena: they were simply lumped together as the

empirical basis for scientific theorizing.

More recent discussions, however, make a threefold distinction between data, phenomena and theory. This started with the work of Bogen and Woodward (1988), who argue that phenomena rather than data are the object of scientists' activities of explanation and prediction, where data can serve as evidence for the existence of phenomena (since they are caused by them). Phenomena, then, are regularities that can be either observable or unobservable (cf. Hacking, 1983). While data are "idiosyncratic to particular experimental contexts", phenomena have "stable, repeatable characteristics which will be detectable by means of a variety of different procedures, which may yield quite different kinds of data" (Bogen and Woodward, 1988, 317).

Bogen and Woodward's view of phenomena is uncompromisingly realist and seems to imply that phenomena are independent of our conceptualizations. Their account has triggered extensive debate and inspired more elaborate analyses of the nature of phenomena. Thus, Hans Radder (2006a), Mieke Boon (2020) and Michela Massimi (2022) have developed accounts with a Kantian flavour, in which conceptual frameworks or 'perspectives' play a crucial role in the conception of phenomena. Hasok Chang (2022, 73–75) defends a similar view on phenomena, arguing that phenomena are "realities, or attributes of realities" that are "mind-framed but not mind-controlled". He rejects the traditional realist idea "that reality has well-defined parts and properties that exist independently of all conceptualization", which he coins the *fallacy of prefiguration*.

The fact that concepts form the basis of scientific theories, and are also crucial for the constitution (or at least the recognition) of phenomena, has important implications for the nature of scientific understanding. As we have seen above in Section 3.1, on De Regt's view scientific understanding of phenomena requires intelligible theories, and hence concepts. This idea aligns with the generally accepted view that concepts are crucial for understanding. Morrison and Morgan, for instance, already stated that it "is th[e] process of interpreting, conceptualising and integrating that goes on in model development which [...] provides the starting point for understanding" (Morrison and Morgan, 1999, 31–3; see also Elgin, 2017, 120). At the same time, the phenomena that are the object of understanding are also partly dependent on conceptualizations.[14] This fundamental role for concepts in the process of achieving scientific understanding of phenomena raises the question of whether, and if so how, human scientists will always be able to understand the results of ML-driven science. For if DNNs acquire concepts (or functional proxies for these; Boge, 2024) that humans have no grasp of, they can perhaps recognize phenomena and develop theories that are unintelligible to humans. In the next sections we will discuss this issue in more detail and consider ways in which the situation might be remedied.

3.3 Paths to the Future (I): Integrating Theory and Machine Learning

A notable response to the challenges raised by ML in PP, as discussed in this chapter, connects to the theory-robustness trade-off mentioned in Sect. 2.2. Recall that, at least for the studies discussed therein, a certain trade-off between theory-freedom and robustness was noted: Training an autoencoder on data contaminated to a small extent with anomalous data (Farina et al., 2020; Heimel et al., 2019) appeared to allow the finding of excesses of such anomalies, as the autoencoder remained unable to successfully reconstruct them after training. However, as shown in the study of Finke et al. (2021), this apparent ability actually depended on the complexity of the 'anomaly' in question. Furthermore, no generally successful method for finding anomalies could be determined, *regardless* of their complexity (Finke et al., 2021), and *without* appeal to some sort of physics-model that *provided* (information about) the (simulated) anomalous data (Fraser et al., 2022).

[14] The concepts on which theories are based need not be the same as those related to phenomena, so there is not necessarily a circularity involved. In particular, the concepts developed by a DNN in order to recognise a novel phenomenon need not be identical to concepts contained in any humanly known theory.

However, from the vantage point of scientific understanding, it seems this vice might be turned into a virtue. Recall that an automated discovery of anomalies, as envisioned by particle physicists, means a *loss* of understanding: According to CUP, we need an intelligible theory T to build models that provide us with understanding of phenomena, where intelligibility may be evidenced by the fact that scientists can recognize qualitatively characteristic consequences of T without performing exact calculations, according to CIT_1. However, as was mentioned already in Sect. 3.1, DNNs do not seem to be the right sorts of models to provide us with understanding. Let us briefly explain this in more detail.

First off, one could be misled to think that DNNs *are* understanding-promoting, as they are constructed under the auspices of an intelligible theory, i.e., deep learning theory (Goodfellow et al., 2014; Shalev-Shwartz and Ben-David, 2014, Ch. 20). Furthermore, successful architectures are often constructed in fairly heuristic ways that are only qualitatively sanctioned by the fundamental theory, rather than corresponding to rigorously proven properties. Hence, following CIT_1, we may take this to show that deep learning theory is intelligible to ML scientists.

However, note that this does not at all establish how DNNs can (and should) themselves promote understanding *about the relevant kind of target*: At best, it shows that, even though many details are presently ill-understood, we have a kind of general grasp of *how DNNs work*. Since their functioning is generally modelled on selective aspects of cognition, neuroscience, and statistical learning, this sort of intelligibility might help us towards an understanding of some aspects of human (or biological) learning – if ever so crudely and selectively (Buckner, 2018; Chirimuuta, 2021). But this, emphatically, doesn't imply that we thereby also obtain an understanding of the *subject matter analysed with* DNNs—in our case: particle physics.

Furthermore, we mentioned the fact that ML 'models' in general are not the sorts of models that mediate between theories and phenomena. Rather, they are similar to purely predictive models, as used, e.g., in descriptive statistics (Boge, 2024). The crucial point to recognize is that DNNs are by and large not *designed to represent* anything about a targeted system or phenomenon: Their elements (weights, biases, activation functions) are not *ab initio* interpreted as referring to a system's states and properties. These elements are actually devoid of any content at the outset and merely employed for the purpose of getting predictions right; something that Boge (2022) has coined being 'instrumental qua devoid of content'. At best, typical ML models are used to, say, represent the statistical distribution of certain features across data, and nothing more.

It is a common, empirically supported assumption, though, that DNNs *develop* representations (or models) during their training (e.g. Buckner, 2018; Goodfellow et al., 2016; López-Rubio, 2020, and above), which then may be understanding-promoting. In fact, we have tacitly bought into this by claiming that DNNs might develop novel concepts, and by making a connection between concepts, theorizing, modelling, and understanding.

However, it is important here to take into account also the specific sort of *opacity* that accompanies DNNs (and possibly other, similarly complex and flexible ML systems). As argued at some length in Boge (2022),[15] it is not only opaque *how* the machine learns (and functions), but also *what* it learns, i.e., what hidden pieces of information *in the data* drive the parameter-updating that we call 'learning'. As the study by Chang et al. (2018) discussed in Sect. 2.1 shows, these pieces of information can be highly relevant for our understanding of the subject matter: Having the mass-information in stock clearly alters the range of models we could employ to render the observed tracks in the detector understandable.

This 'what-opacity' (Boge, 2022), conjoined with the specific instrumental character of ML models in PP, implies that it will usually require additional effort to get from the recognition

[15] Beisbart (ch. 1, this volume) offers an account of opacity that is phrased in different terms but roughly consistent with Boge's; Formanek (ch. 2, this volume) instead offers a deflationary take on opactiy, though he does not engage with Boge's arguments in detail.

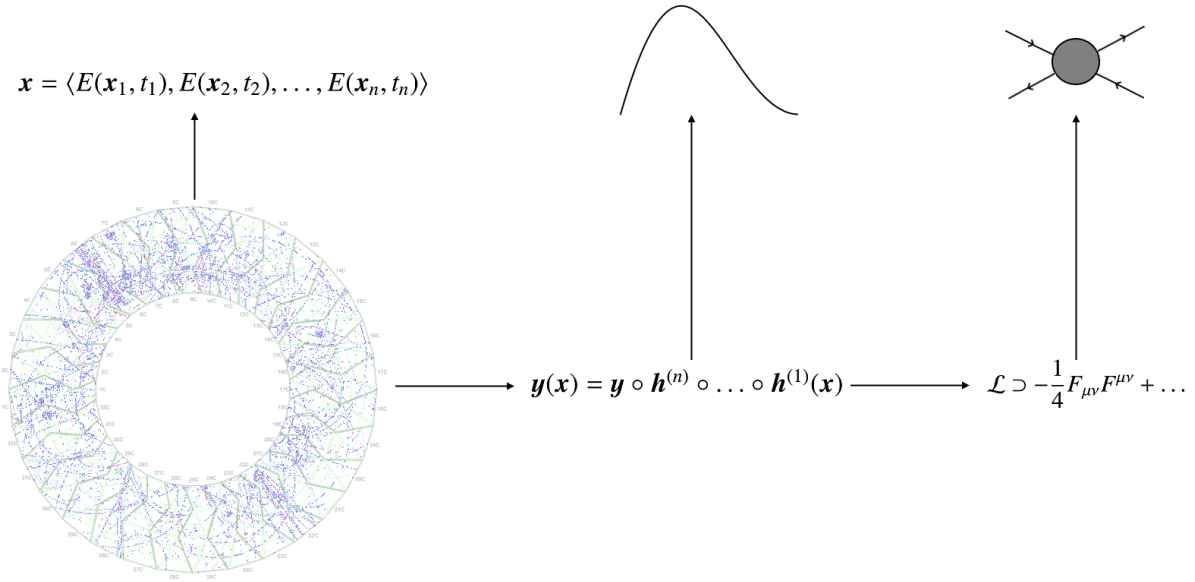

Figure 5: Modeling steps that can lead from ML successes to scientific understanding. The first step means interpreting certain data, which are just currents received from a detector, in terms of energy-deposits from particle-interactions. The second step means (e.g.) running a DNN-classifier that distinguishes signal- from background-events on this data and interpreting it as providing a statistical relation, such as a likelihood ratio. The third step then means connecting this likelihood ratio back to explanatory physics-models, which might be a formidable task if the underlying processes are entirely unknown. Taken from Boge (2022) under a CC BY 4.0 license. Color available online.

of what the ML model itself is used to represent (say, a statistical ratio or distribution) to an understanding of the subject matter: The model is *used* to represents fairly little (only said distribution or ratio), and what it might 'represent *to itself*' will be difficult to identify.[16] In particular, connecting a successful DNN up with more understanding-promoting representations, such as theory-bound physics models, will require additional modelling. (see Fig. 5).

We can now easily see how the vice of theory-presence in ML-based discovery-methods might become a virtue regarding understanding: If theoretical knowledge enters at various stages, this additional physics-modelling may profit from knowledge about how theory is used to delimit the things the DNN can 'learn' in the first place.

Approaches which have this idea built into them go by the name of 'hybrid ML'. Various distinct proposals have been subsumed under this label, such as patching together theory-inspired and ML models or replacing parts of DNNs by known, theoretically motivated mappings (Karpatne et al., 2017; Maier et al., 2022; Reichstein et al., 2019).[17] Of greatest interest to us here are theory-based approaches to 'inductive bias'.

Broadly speaking, this means "the incorporation of prior knowledge that biases the learning mechanism." (Shalev-Shwartz and Ben-David, 2014, 3) A little more specifically, it means restricting the class of functions that can be instantiated by the given DNN or ML system (ibid., 16). As has been pointed out by Sterkenburg and Grünwald (2021), it is possible to interpret the so called 'no free lunch theorems' (Wolpert, 1996), which say that under certain modest-seeming assumptions, "every learning algorithm can be expected to do no better (or worse) than random guessing" (Sterkenburg and Grünwald, 2021, 9982), as suggesting that inductive bias is *needed*

[16] We will turn to a specific class of proposals on how this might be done below.
[17] For an appraisal in the context of astronomy and cosmology, see also Meskhidze (ch. 17, this volume).

to make ML work.

Insofar as this interpretation is correct, there is hardly anything puzzling about the above findings on autoencoders and their model-dependence:[18] Some idea about what a given anomaly looks like is needed; otherwise we don't stand a chance of *reliably* detecting it.

Furthermore, from the vantage point of scientific understanding, the situation is as follows: Insofar as a given implementation of inductive bias is motivated by theory – say, using a specific loss-function whose shape is guided by a physics model (Fraser et al., 2022) –, we can conclude that a successful autoencoder trained in such a way will find signals consistent with our theoretical expectations. This immediately allows us to understand the phenomena discovered, but of course it won't be anomaly-detection anymore: This would require the discovery of phenomena *inconsistent* with any theoretical expectations.

One might try to mediate between this loss in theory-freedom and gain robustness by employing an extended scheme that uses a broad range of differently trained autoencoders and thus looks for signals in the data consistent with various theoretical scenarios. This would certainly maintain a respectable increase in discovery potential by means of ML while also making the gap to theoretical understanding smaller. But, admittedly, it would certainly also mean losing our handle on automated, largely theory-free discovery of phenomena.

3.4 Paths to the Future (II): Learning from the Machine?

Above we have seen that hybrid approaches, combining data-driven methods with theory-driven ones, can improve robustness and at the same time allow (at least in principle) for scientific understanding in the sense of CUP. However, the question remains whether novel conceptualizations developed by DNNs, that form the basis of their modeling and theorizing, can be made intelligible to human scientists. In this section we address this question by discussing two recent proposals for transferring understanding generated by AI, esp. DNNs, to humans.

In their paper 'On scientific understanding with artifical intelligence' Krenn et al. (2022) analyze how AI may contribute to achieving scientific understanding, by reviewing the current use of AI in science and discussing future prospects. They submit that AI might foster scientific understanding in three ways: (1) as a "computational microscope"; (2) as a "resource of inspiration"; and (3) as an "agent of understanding". (1) comprises uncovering patterns in complex data sets, and is already widely used. (2) concerns identifying surprises in the data or in the scientific literature, which may require that novel concepts are developed. Currently, interpretation and conceptualization has to be done by human scientists, but Krenn et al. (2022, 754-756) do not preclude the possibility that future AI will be able to come up with novel concepts by itself . The third case goes one step further. Here AI generates new explanations and has thereby acquired novel understanding on its own: the AI is an "agent of understanding". Also this future is still ahead of us but not unlikely according to Krenn et al. In light of what we have discussed above, development of novel concepts and of novel explanations and understanding is closely related, and in any case raises the question of whether, and if so how, human scientists can learn from AI. If the newly developed concepts and understanding are not intelligible to humans, the result is a divide between human and artificial scientific understanding, which would be an obstacle to making progress in scientific understanding as a whole.

So, the crucial question is whether, and if so how, the AI can transfer its understanding to humans. To answer this question, we start with a suggestion made by Krenn et al.: they propose a 'scientific understanding test' for determining whether an AI has scientific understanding. They start with De Regt's above-mentioned criterion for intelligibility CIT_1 and suggest the following analogous condition: "An AI gained scientific understanding if it can recognize qualitatively

[18] Note that this is meant in a purely intuitive sense: Autoencoders trained as described above instantiate unsupervised learning, so they are not subject to the theorems mentioned here.

characteristic consequences of a theory without performing exact computations and use them in a new context" (Krenn et al., 2022, 767). By itself this is not yet sufficient for actually assessing whether an AI has understanding. Therefore, Krenn et al. add a second condition, according to which "an AI gained scientific understanding if it can transfer its understanding to a human expert". Combining the two conditions, they formulate their 'scientific understanding test' for judging whether an AI has in fact gained understanding:

> A human (the student) interacts with a teacher, either a human or an artificial scientist. The teacher's goal is to explain a scientific theory and its qualitative, characteristic consequences to the student. Another human (the referee) tests both the student and the teacher independently. If the referee cannot distinguish between the qualities of their non-trivial explanations in various contexts, we argue that the teacher has scientific understanding. (Krenn et al., 2022, 767)

This obviously resembles a Turing test (as Krenn et al. readily admit). However, rather than using it as such, it can also serve a different purpose, namely as providing a link between human and artificial understanding and as a means to examine the extent to which humans can learn from AI. But the test does not yet tell us *how exactly* the referee measures the quality of the alleged understanding transfer from teacher to student. Krenn et al. do not elaborate on this issue, but Barman et al. (2024) provide an answer. They argue that evaluation of understanding in both humans and other agents (e.g. AIs) should be based on their *abilities* to perform relevant tasks, an approach that emphasizes the ability-aspect of concepts. Barman et al. present a general framework *measuring* an agent's scientific understanding:

AUP The degree to which agent A scientifically understands phenomenon P can be determined by assessing the extent to which (i) A has a sufficiently complete representation of P; (ii) A can generate internally consistent and empirically adequate explanations of P; (iii) A can establish a broad range of relevant, correct counterfactual inferences regarding P. (Barman et al., 2024)

The idea behind this proposal is that scientific understanding is not an all-or-nothing affair (captured by necessary and sufficient conditions) but a matter of degree, comprising three levels of increasing value. The framework allows for different ways to measure the degree to which A has (i), (ii) and/or (iii). Barman et al. focus on Large Language Models (LLMs), for which they suggest the following specification: "[AUP](i-iii) can be measured, given a certain context (series of prompts) via what-, why-, and what-if-questions respectively." The ability to correctly answer what-questions reflects possession of relevant information about P (i), while an ability to answer why-questions regarding P reflects the capability to produce an explanation of it (ii). Finally, the ability to answer what-if-questions reflects competency at establishing counterfactual inferences about P (iii). This requires not only having a good explanation but also knowing how to use it, and can accordingly be linked to an agents' breadth and depth of understanding. Barman et al. (2024) argue that application of this approach to the scientific understanding test of Krenn et al. (2022) allows for measuring the increase of understanding in the student, and hence the understanding transfer from AI to humans. Extending this approach to AI in general would require integration of LLMs with other types of DNNs, a goal the pursuit of which is currently still in its infancy (e.g. Singh et al., 2022) but will plausibly be realized in the near future.

4 Conclusions

In this chapter we have explored the possibility that DNNs in particle physics can (I) autonomously discover novel, physically meaningful concepts, and (II) recognize anomalous phenomena that would otherwise be concealed from human sight. Both (I) and (II) would prima

facie create obstacles for gaining understanding for the following reasons. First, concepts are the building blocks of theories that are required for understanding, and hence the unintelligibility of ML-generated concepts entails a failure of theoretical understanding. Second, anomalies by definition constitute ill-understood phenomena whose resolution would require extraordinary effort in the case of ML-based discovery.

As regards (I), we have argued that it is not implausible that DNNs can possess concepts (or at least be able to mimic concept-possession) and perhaps also develop novel concepts. However, it is not guaranteed that such novel concepts are intelligible to humans, and if they are not, humans cannot use them directly for acquiring scientific understanding. We have suggested that this problem might be resolved by invoking two recent proposals for establishing and measuring a transfer of scientific understanding between artificial agents, such as DNNs, and humans.

As it turns out, the impression that (II) is indeed the case is deceptive. Existing methods require theoretical input to become robust, that is, input on what the 'anomalies' to be discovered look like, so these would actually not be anomalies anymore. From the vantage-point of understanding, this seems like a virtue though, as the gap between discovered phenomena and our understanding of these is immediately reduced or even closed: If we know the potential new phenomena that a DNN can recognize through theories, we may already understand them fairly well. Disappointingly, this also means that the discovery-potential associated with DNNs in PP is currently less impressive than one might have initially thought.

Acknowledgements We thank Helen Meskhidze for a number of comments. FJB acknowledges generous funding by the German Research Foundation (DFG), for his Emmy Noether grant *UDNN: Scientific Understanding and Deep Neural Networks* (grant number 508844757), as well as various discussions with particle physicist Michael Krämer from RWTH Aachen University.

Biographical Information Florian J. Boge is Junior-Professor for Philosophy of Science with a focus on Artificial Intelligence at TU Dortmund University and leader of the DFG-funded Emmy Noether research group *UDNN: Scientific Understanding and Deep Neural Networks*. He holds a PhD from the University of Cologne, for a thesis on the interpretation of quantum physics. Boge is associate editor for the *European Journal for Philosophy of Science* and was a postdoc in the subproject *The Impact of Computer Simulations and Machine Learning on the Epistemic Status of LHC Data* of the interdisciplinary DFG research unit *The Epistemology of the Large Hadron Collider*. He is also an associated PI of the Lamarr Institute for Machine Learning and Artificial Intelligence.

Henk W. de Regt is Professor of Philosophy of Natural Sciences at the Institute for Science in Society, Radboud University, The Netherlands. He holds an MSc in foundations of physics (Utrecht University) and a PhD in philosophy (Vrije Universiteit Amsterdam). His research centers on scientific understanding and explanation, which he has recently extended to questions regarding the role of AI in science. His monograph *Understanding Scientific Understanding* (Oxford University Press, 2017) won the 2019 Lakatos Award. De Regt co-founded the *Society for Philosophy of Science in Practice* (SPSP) and the *European Society for the Philosophy of Science* (EPSA).

References

Albertsson, K., Altoe, P., Anderson, D., Andrews, M., Espinosa, J. P. A., Aurisano, A., Basara, L., Bevan, A., Bhimji, W., Bonacorsi, D., et al. (2018). Machine learning in high energy physics community white paper. *Journal of Physics: Conference Series*, 1085(2):022008.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. (2019). Strike

(with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854.

Ananthaswamy, A. (2021). AI designs quantum physics experiments beyond what any human has conceived. *Scientific American*, July 2. https://www.scientificamerican.com/article/ai-designs-quantum-physics-experiments-beyond-what-any-human-has-conceived/.

Arabatzis, T. (2019). What are scientific concepts? In McCain, K. and Kampourakis, K., editors, *What is scientific knowledge?*, pages 85–99. Routledge.

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

Barman, K. G., Caron, S., Claassen, T., and de Regt, H. (2024). Towards a Benchmark for Scientific Understanding in Humans and Machines. *Minds & Machines*, 34:6.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. *arXiv preprint arXiv:1704.05796*.

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

Bechtle, P., Chall, C., King, M., Krämer, M., Mättig, M., and Stöltzner, M. (2022). Bottoms up: The standard model effective field theory from a model perspective. *Studies in History and Philosophy of Modern Physics*. preprint: https://arxiv.org/pdf/2201.08819.pdf.

Beisbart, C. (MS). In which ways is machine learning opaque? In Pozzi, G. and Durán, J. M., editors, *Philosophy of Science for Machine Learning*. Springer.

Boge, F.J., Poznic, M. (2021). Machine Learning and the Future of Scientific Explanation. *Journal for General Philosophy of Science*, 52:171–176.

Boge, F., Grünke, P., and Hillerbrand, R. (2022). Minds and machines special issue: Machine learning: Prediction without explanation? *Minds & Machines*, 32:1–9.

Boge, F. J. (2021). Why trust a simulation? models, parameters, and robustness in simulation-infected experiments. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/716542.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, 32(1):43–75.

Boge, F. J. (2024). Functional concept proxies and the actually smart hans problem: What's special about deep neural networks in science. *Synthese*, 203(16):39 pp. https://doi.org/10.1007/s11229-023-04440-8.

Boge, F. J., Krämer, M., and Zeitnitz, C. (MS). Deep learning robustness for scientific discovery: The case of anomaly detection. *unpublished*.

Boge, F. J. and Zeitnitz, C. (2021). Polycratic hierarchies and networks: what simulation-modeling at the lhc can teach us about the epistemology of simulation. *Synthese*, 199(1-2):445–480.

Bogen, J. and Woodward, J. (1988). Saving the phenomena. *The Philosophical Review*, XCVII(3):303–352.

Boon, M. (2020). The role of disciplinary perspectives in an epistemology of scientific models. *European Journal for Philosophy of Science*, 10:article 31.

Buchholz, O. (2023). The deep neural network approach to the reference class problem. *Synthese*, 201: 111.

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

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

Buijsman, S. (MS). Machine learning models as mathematics. In Durán, J. M. and Pozzi, G., editors, *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Synthese Library. Springer.

Butterworth, J. M., Grellscheid, D., Krämer, M., Sarrazin, B., and Yallup, D. (2017). Constraining new physics with collider measurements of Standard Model signatures. *Journal of High Energy Physics*, 2017(3):78.

Carey, S. (2011). *The Origin of Concepts*. Oxford series in cognitive development. Oxford University Press.

Cartwright, N. (1983). *How the laws of physics lie*. Oxford University Press, Oxford.

Chang, H. (2022). *Realism for realistic people: A new pragmatist philosophy of science*. Cambridge University Press, Cambridge.

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

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, 199(1-2):767–790.

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

CMS Collaboration (2014). Observation of the diphoton decay of the Higgs boson and measurement of its properties. *European Physical Journal C*, 74, 3076. https://doi.org/10.1140/epjc/s10052-014-3076-z

Dellsén, F. (2021). Understanding scientific progress: the noetic account. *Synthese*, 199:11249–11278.

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

de Regt, H. W. (2020). Understanding, values, and the aims of science. *Philosophy of Science*, 87:921–932.

de Regt, H. W., Leonelli, S., and Eigner, K. (2009). *Scientific understanding: Philosophical perspectives*. University of Pittsburgh Press.

Delfino, M. (2020). Distributed computing. In Fabjan, C. W. and Schopper, H., editors, *Particle Physics Reference Library, Volume 2*, pages 613–644. Cham: Springer.

Dillon, B. M., Favaro, L., Plehn, T., Sorrenson, P., and Krämer, M. (2022). A normalized autoencoder for lhc triggers. *ArXiv:2206.14225*, [hep-ph]. https://doi.org/10.48550/arXiv.2206.14225.

Duede, E. (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, page 1–13. DOI=10.1017/psa.2023.8.

Elgin, C. Z. (2017). *True Enough*. Cambridge MA, London: MIT Press.

Facchini, A. and Termine, A. (2021). Towards a taxonomy for the opacity of ai systems. In Müller, V. C., editor, *Conference on Philosophy and Theory of Artificial Intelligence*, pages 73–89. Springer.

Farina, M., Nakai, Y., and Shih, D. (2020). Searching for new physics with deep autoencoders. *Physical Review D*, 101(7):075021.

Finke, T., Krämer, M., Morandini, A., Mück, A., and Oleksiyuk, I. (2021). Autoencoders for unsupervised anomaly detection in high energy physics. *Journal of High Energy Physics*, 2021(6):161.

Fodor, J. (1994). Concepts: A potboiler. *Cognition*, 50(1-3):95–113.

Formanek, N. (MS). How i stopped worrying and learned to love opacity. In Pozzi, G. and Durán, J. M., editors, *Philosophy of Science for Machine Learning*. Springer.

Fraser, K., Homiller, S., Mishra, R. K., Ostdiek, B., and Schwartz, M. D. (2022). Challenges for unsupervised anomaly detection in particle physics. *Journal of High Energy Physics*, 2022(3):1–31.

Freiesleben, T. (MS). Artificial neural nets and the representation of human concepts. In Durán, J. M. and Pozzi, G., editors, *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Synthese Library. Springer.

Giere, R. (2006). *Scientific Perspectivism*. University of Chicago Press.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge (MA), London: The MIT Press.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, Massachusetts: Harvard Univeristy Press.

Greif, H. (2022). Analogue models and universal machines. paradigms of epistemic transparency in artificial intelligence. *Minds and Machines*, 32(1):111–133.

Grimm, S. R., Baumberger, C., and Ammon, S. (2017). *Explaining understanding*. Routledge.

Hacking, I. (1983). *Representing and Intervening*. Cambridge, New York: Cambridge University Press.

Hampton, J. A. (2006). Concepts as prototypes. *Psychology of learning and motivation*, 46:79–113.

Heimel, T., Kasieczka, G., Plehn, T., and Thompson, J. M. (2019). Qcd or what? *SciPost Physics*, 6(030):21.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*.

Iten, R., Metger, T., Wilming, H., Del Rio, L., and Renner, R. (2020). Discovering physical concepts with neural networks. *Physical Review Letters*, 124(1):010508.

Karaca, K. (2018). Lessons from the large hadron collider for model-based experimentation. *Synthese*, 195(12):5431–5452.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, 29(10):2318–2331.

Kästner, L. and Crook, B. (MS). Don't fear the bogeyman: On why there is no prediction-understanding trade-off for deep learning in neuroscience. In Pozzi, G. and Durán, J. M., editors, *Philosophy of Science for Machine Learning*. Springer. online: http://philsci-archive.pitt.edu/22344/.

Kieval, P. H. (MS). Representation learning without representationalism: A non-representational account of deep learning models in scientific practice. In Durán, J. M. and Pozzi, G., editors, *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Synthese Library. Springer.

Koningsveld, H. (1973). *Empirical Laws, Regularity and Necessity*. Wageningen: H. Veenman & Zonen B.V.

Krenn, M., Pollice, R., Guo, S. Y., Aldeghi, M., Cervera-Lierta, A., Friederich, P., dos Passos Gomes, G., Häse, F., Jinich, A., Nigam, A., et al. (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, 4(12):761–769.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*. University of Chicago Press, second, enlarged edition.

Lakatos, I. (1970). History of science and its rational reconstructions. *Proceedings of the biennial meeting of the philosophy of science association*, 1970:91–136.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In Harding, S. G., editor, *Can Theories be Refuted?*, page 205–259. D. Reidel Publishing Company.

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.

Maier, A., Köstler, H., Heisig, M., Krauss, P., and Yang, S. H. (2022). Known operator learning and hybrid machine learning in medical imaging—a review of the past, the present, and the future. *Progress in Biomedical Engineering*, 4(2):022002.

Margolis, E. and Laurence, S. (2023). Concepts. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2023 edition.

Martens, N. C. M. and Lehmkuhl, D. (2020). Dark matter = modified gravity? scrutinising the spacetime–matter distinction through the modified gravity/ dark matter lens. *Studies in History and Philosophy of Modern Physics*, 72:237–250.

Massimi, M. (2022). *Perspectival realism*. Oxford University Press, Oxford.

Mattia, F. D., Galeone, P., Simoni, M. D., and Ghelfi, E. (2021). A survey on gans for anomaly detection. *arXiv*, cs.LG(1906.11632).

McCoy, C. D. and Massimi, M. (2018). Simplified models: a different perspective on models as mediators. *European Journal for Philosophy of Science*, 8(1):99–123.

Meskhidze, H. (MS). Beyond classification and prediction: The promise of physics-informed machine learning in astronomy and cosmology. In Pozzi, G. and Durán, J. M., editors, *Philosophy of Science for Machine Learning*. Springer.

Morrison, M. and Morgan, M. S. (1999). Models as mediating instruments. In Morrison, M. and Morgan, M. S., editors, *Models as Mediators*, pages 10–37. Cambridge University Press.

Páez, A. (MS). Axe the x in xai: A plea for understandable ai. In Durán, J. M. and Pozzi, G., editors, *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Synthese Library. Springer.

Passon, O. (2019). On the interpretation of Feynman diagrams, or, did the LHC experiments observe $h \to \gamma\gamma$? *European Journal for Philosophy of Science*, 9(2):20.

Piccinini, G. (2011). Two kinds of concept: Implicit and explicit. *Dialogue*, 50(1):179–193.

Quine, W. (1969). Natural kinds. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel*, pages 5–23. Dordrecht: Springer.

Radder, H. (2006a). *the world observed / the world conceived*. University of Pittsburgh Press.

Radder, H. (2006b). *The world observed/the world conceived*. University of Pittsburgh Press.

Räz, T. (2023). Methods for identifying emergent concepts in deep neural networks. *Patterns*, 4(6).

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat, f. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204.

Rosaler, J. and Harlander, R. (2019). Naturalness, wilsonian renormalization, and 'fundamental parameters' in quantum field theory. *Studies in History and Philosophy of Modern Physics*, 66:118–134.

Rowbottom, Darrell P. (2023). *Scientific Progress*. Cambridge: Cambridge University Press.

Schwartz, M. (2014). *Quantum Field Theory and the Standard Model*. Cambridge University Press.

Schwartz, M. D. (2021). Modern machine learning and particle physics. *arXiv preprint arXiv:2103.12226*.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Singh, C., Morris, J. X., Aneja, J., Rush, A. M., and Gao, J. (2022). Explaining patterns in data with language models via interpretable autoprompting. *arXiv preprint arXiv:2210.01848*.

Sterkenburg, T. F. and Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese*, 199(3-4):9979–10015.

Sullivan, E. and Kasirzadeh, A. (MS). Explanation hacking: The perils of algorithmic recourse. In Durán, J. M. and Pozzi, G., editors, *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Synthese Library. Springer.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

Zhang, H., Chen, H., Song, Z., Boning, D., Dhillon, I. S., and Hsieh, C.-J. (2019). The limitations of adversarial training and the blind-spot attack. *arXiv preprint arXiv:1901.04684*.