Synthese (2024) 203:16
https://doi.org/10.1007/s11229-023-04440-8

ORIGINAL RESEARCH

Functional Concept Proxies and the Actually Smart Hans Problem: What's Special About Deep Neural Networks in Science

Florian J. Boge[1,2]

Received: 23 November 2021 / Accepted: 23 November 2023

Abstract

Deep Neural Networks (DNNs) are becoming increasingly important as scientific tools, as they excel in various scientific applications beyond what was considered possible. Yet from a certain vantage point, they are nothing but parametrized functions $f_{\theta}(x)$ of some data vector x, and their 'learning' is nothing but an iterative, algorithmic fitting of the parameters to data. Hence, what could be special about them as a scientific tool or model? I will here suggest an integrated perspective that mediates between extremes, by arguing that what makes DNNs in science special is their ability to develop *functional concept proxies* (FCPs): Substructures that occasionally provide them with abilities that correspond to those facilitated by concepts in human reasoning. Furthermore, I will argue that this introduces a problem that has so far barely been recognized by practitioners and philosophers alike: That DNNs may succeed on some vast and unwieldy data sets because they develop FCPs for features that are not transparent to human researchers. The resulting breach between scientific success and human understanding I call the 'Actually Smart Hans Problem'.

Keywords Deep Neural Networks · Concepts · Reasoning · Clever Hans Problem · Automated science

> *It is very difficult for us to deconstruct a neural network to figure out exactly what concepts the algorithm is "learning" [...]. In other words, AlphaFold has improved our* ability *to predict a protein structure from its sequence; but hasn't directly increased our* understanding *of how protein sequence relates to structure.*
>
> *—Foldit staff member 'bkoep' (https://fold.it/portal/node/2008706, posted January 31st, 2020; orig. emph.)*

✉ Florian J. Boge
florian-johannes.boge@udo.edu

1 Interdisciplinary Centre for Science and Technology Studies (IZWT), Wuppertal University, Wuppertal, Germany

2 Institute for Philosophy and Political Science (IPP), TU Dortmund, Emil-Figge-Str. 50, room 2.247, 44227 Dortmund, Germany

Published online: 27 December 2023

AI systems can learn to identify patterns, but they cannot understand the concepts behind those patterns.

—GPT3, when given the prompt "Write an essay proving that an AI system trained on form can never learn semantic meaning" (https://scottaaronson.blog/, posted April 24th, 2022)

1 Introduction

Without a doubt, Deep Neural Networks (DNNs) have become increasingly important as scientific tools, as they excel in various scientific applications beyond what was considered possible. Nevertheless, there is a strong continuity between present-day DNNs and traditional data analysis methods and a general sense that there may really be nothing new here, as reflected by the famous 'internet meme' displayed in Fig. 1.

A stark example of this is Google's 'AlphaFold2', which vastly outperformed 100 rival methods in 2020's Critical Assessment of Structure Prediction (CASP14). Predicting protein structures from amino acid sequences has been a hard problem for decades (e.g. Branden & Tooze, 1999). But in 2/3 of the test cases in CASP14, AlphaFold2 predicted structures to within the experimental accuracy of their empirically determined shapes, and came close in the remaining cases. Because of this impressive leap ahead, AlphaFold2 has been hailed an outright 'game changer' (see Callaway, 2020).

Google's DeepMind team (Jumper, 2021a, b) used the novel 'Transformer' algorithm (Vaswani et al., 2017), originally developed for natural language processing, in it's 'trunk'. Furthermore, unlike its already successful predecessor (Senior et al., 2020), AlphaFold2 integrated information on the evolutionary history of proteins, the known physical driving forces pertaining to molecules, and geometric information to constrain the possible protein structures. Still, the bioinformatics community found nothing *fundamentally* new in this approach:

> In some respects, seeing the final complete description of the method was a tiny bit disappointing, after the huge anticipation that had built up following the CASP14 meeting. [...] In many respects, AlphaFold2 is 'just' a very well-engineered system that takes many of the recently explored ideas in the field, such as methods to interpret amino acid covariation, and splices them together seamlessly using attention processing. (Jones & Thornton, 2022, p. 16)

In detail, one can map pretty much all success-driving elements of AlphaFold2 to well-known principles of traditional Machine Learning (ML): The arrangement of amino acid sequences into data matrices that reveal evolutionary connections between proteins is nothing but "a separate pre-processing step" (Petti et al., 2021, p. 2); i.e., something "that involves transforming raw data into an understandable format" (Mariani et al., 2021, p. 109). The Transformer algorithm computes a non-linear function of dot-products between linearly transformed vectors from these matrices. Informally, this 'contextualizes' each vector in the sense of acknowledging the importance of other vectors surrounding it. More formally, we have a combination of linear and non-linear functions with learnable weights, so 'just' a specific DNN architecture.

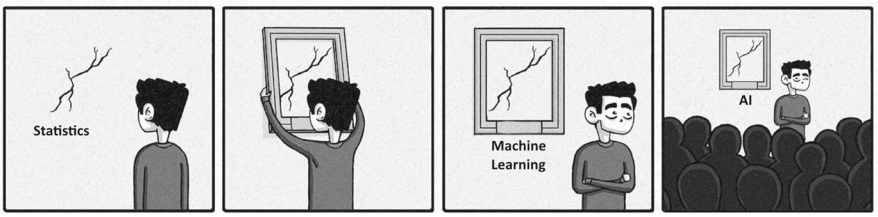

Fig. 1 Famous 'internet meme' displaying the relation between statistics, Machine Learning and Artificial Intelligence. Original image by SandSerif, courtesy of the artist

Finally, in AlphaFold2's 'structure module' (a second, subsequent DNN), physico-geometric information strongly influences the space of potential outputs. But this is nothing but an *inductive bias*[1]; something well-known from statistical learning theory (Shalev-Shwartz & Ben-David, 2014, p. 16) and arguably necessary for real-world success (Sterkenburg & Grünwald, 2021).

Another example is Google's 'LaMDa' chatbot—an equally Transformer-based, state of the art DNN whose extraordinary skill in describing its alleged feelings and emotions convinced at least one Google engineer of its sentience. This sentiment was, however, met with a lot of criticism by large chunks of the Artificial Intelligence (AI) community, and ultimately led to said engineer's suspension.[2] As Alberto Delgado from the National University of Columbia comments on his twitter account,

> It is mystical to hope for awareness, understanding, common sense, from symbols and *data processing using parametric functions in higher dimensions*.[3]

So what could possibly be *special* about even such advanced DNNs as AlphaFold2 and LaMDa? Are they not just fancy, complex, generic data models, subject to an iterative, parametric statistical optimization? In a way, I believe the answer is 'yes'. But it is vital to notice that this does not imply that there is nothing remarkable about these ML systems, especially when they are used to do science. To illustrate the point a bit, consider how the human body is biochemically speaking just a hypercomplex macro-molecule. Yet human bodies contain brains that provide them with abilities way beyond what less complex molecules can do. Without intending too close an analogy between brains and DNNs here, this readily illustrates how complexity, rather than type-identity, could possibly be relevant for the presence of special features.

In this paper, I will hence suggest an integrated perspective that mediates between extremes, by focusing on a particular feature that makes them appear like cognitive agents: Their apparent ability to acquire *concepts*, and that this seems to be a major reason for their success (Buckner, 2018; López-Rubio, 2020). On the other hand, it seems safe to say that there is little evidence that present DNNs are *conscious*, and

[1] See https://predictioncenter.org/casp15/doc/presentations/2020_12_01_TS_predictor_AlphaFold2.pdf, slide 7 (checked 08/22).

[2] E.g. https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine (checked 07/22).

[3] See https://twitter.com/jadelgador/status/1535979040925958144?s=21&t=ncEQgno59MKl5OR6KJbz1A (checked 07/22; emphasis added).

some might think this is reason enough to doubt the possession of concepts. More importantly, the *kind of mistakes* DNNs typically make, together with the decisively semantic properties, plasticity and systematicity of concepts, should promote some healthy skepticism here.[4] I will hence argue that they do not possess concepts but *functional concept proxies* (FCPs)[5] Substructures that provide them with abilities that correspond to, or exceed, those facilitated by concepts in human reasoning, but fail to do so under certain circumstances, and in ways that exhibit their mere proxyhood. I shall offer more precise definitions in Sect. 2.

Now I said that FCPs are the major reason for DNNs' success, but they need not always promote success, or at least not all across the board: They could possibly also be misguided. In particular, this may happen when a DNN learns to specialize to extremely subtle features of the data that do not generalize well beyond the training and testing data sets. The ensuing problem resulting from this is known as the *Clever Hans Problem* (CHP) in the technical literature.

Given that FCPs are at least often responsible for success, though, and that we use DNNs to process often vast and unwieldy data, I believe there is also an opposite, 'Actually Smart Hans Problem' (ASHP). Imagine the following situation: A DNN is trained on a vast and poorly understood data set. It thereby learns to process, classify, and predict based on well-generalising features of the data that are not easily (if at all) recognizable for humans. Furthermore, no human being has a relevant concept to comprehend, or even identify, these features (yet). Would we not consider this a problem, and find the DNN to have actually outsmarted us by means of the FCPs it has, apparently, developed? Furthermore, given that the DNN does not even need 'real' concepts to so outsmart us, it would likely be unable to sensibly communicate its findings to us. Would we not consider this an uncomfortable situation, in which successful prediction has outpaced scientific understanding?

The paper is essentially organized into two parts: The first one establishes the notion of FCPs and the reasons for embracing it. The second part then applies this notion to scientific examples and outlines how a serious, novel problem may be generated by DNNs' ability to develop (new) FCPs somewhat autonomously. Thus, the second part establishes why philosophers of science should care. The main contribution of the paper, hence, lies in (a) finding the right vocabulary to discuss these subtle issues and (b) establishing (or at least deepening) important connections between the cognitive science-oriented parts of the debate on AI and DNNs and the philosophy of science-oriented one.

As a small caveat, I shall note here that all this may pertain, strictly speaking, not *only* to DNNs, but more generally to ML systems that combine complex learning and very general parameterized mappings in the right ways; DNNs currently being the most prominent sort of such systems.[6] I shall largely suppress this issue below though, as I believe it alters nothing about the philosophical substance of the paper.

[4] See, e.g., (Marcus & Davis, 2020) for similar skepticism, based on similar observations.

[5] Note that I will not offer a principled argument that *all*: DNNs *inevitably* develop (at least) FCPs, though I think the cited evidence makes this plausible.

[6] I have in mind here such things as, say, the automated topological graph-optimizer 'Theseus' used by Krenn et al. (2021) to find new quantum optics experiments.

2 Functional Concept Proxies

2.1 Concepts

In order to say what a proxy for a concept is, I first need to say what I mean by 'concept'. This is less straightforward than it might seem, as extant theories differ among philosophers as well as between disciplines (see Camp, 2009; Machery, 2009). Furthermore, as several anonymous referees have pointed out, given a sufficiently rich notion of concepts, it would fairly straightforward to establish that a DNN cannot, in fact, have concepts. However, this would pretty much mean excluding concept-possession in DNNs by *fiat*, and so it is certainly more interesting to ask whether the same is also true under a fairly modest reading of 'concept'. Appealing to a modest notion at the same time avoids several thorny issues in the philosophy of mind. I will hence appeal to such a modest account, by building on certain reasonably modest criteria for concept-attribution that have been distilled by a number of authors.[7]

Consider Camp's (2009) approach to concepts. Camp compares the notion traditionally employed by philosophers from at least Descartes on, which assumes a strong connection to linguistic capabilities, with psychologists' usage of the word, which is far more permissive. The latter notion, in particular, allows for animals that systematically respond to different stimuli in adequate ways to have concepts. Prima facie, there appears to be stark disagreement between both notions. But Camp (2009, p. 276) argues, the core element that connects both approaches is "an important sort of systematicity."

An example given by her (280 ff.) is the imagined ability of dog D to treat another dog, M, at times as a hunting partner, at times as a threat. This behavior of D towards M might still be different from D's behavior towards yet another dog, N, who was always treated as a threat by D. It hence seems plausible that D has distinct concepts of M and N, as well as distinct concepts HUNTING PARTNER and THREAT. Crucially, the things to be combined are representations of particulars and their ways of being, and the latter representations can be combined in different ways with the former ones.

Camp (2009, p. 276; emph. added) considers a view she calls 'minimalism' about concepts, which has "*any* representational abilities that can be systematically recombined" be conceptual abilities. However, minimalism might be just too minimal, as concept-possession additionally requires *stimulus-independence*:

> it is now extremely well-established that creatures with no more than basic cognition are not confined to *representing only states of affairs that they take themselves to be directly confronting*. [...] a wide range of animals can represent properties at distant locations, and navigate to those locations by novel routes to satisfy their desires [...]. (Camp, 2009, p. 289, emph. added)

Another account compatible with these considerations is that of Newen and Bartels (2007), which builds on the behavior of Parrot 'Alex', studied by Pepperberg (1999). In order to determine whether Alex could be said to have concepts of colors and shapes, Pepperberg designed a number of tests in which Alex had to respond to questions

[7] I owe thanks to an anonymous referee and to Albert Newen for pointing me to relevant references here.

targeting sameness and differences between visual stimuli in particular respects, such as shape and color, number or object type. Alex also had to perform these tasks on never before encountered items or pairs thereof, including sameness and difference-tasks w.r.t. colors not encountered in the test before (cf. Pepperberg, 1999, pp. 58–68).

Following Newen and Bartels (2007, pp. 293–294; emph. added), Alex's success in these tasks nurtures the intuition that "in order to have one concept you should have a *minimal semantic net* including that concept", meaning a system of representations that allows one to identify and re-identify objects and their properties, with representations being stimulus independent and involving some amount of abstraction. 'Abstraction' is here cashed out as going beyond the mere generalization of stimuli into perceptual equivalence classes (such as the presence of a beak for bird-identification; cf. ibid., pp. 292, 295).

Thus, in order to have a concept of a particular color, there must be a concept of a different, contrasting color as well as concepts of at least two further, contrasting properties (such as two distinct shapes), combinable with the former ones (but not one another). These pairs of propertied may be said to lie along different dimensions (cf. ibid., pp. 293–294). Thus, the semantic aspect of concepts is intimately linked with their systematicity and concerns the carrier's ability to form different abstract property-representations and the ways in which they can (and cannot) be combined. These criteria for the presence of a 'minimal semantic net' Newen and Bartels hold to be "satisfied if the behavior of a cognitive system can be *explained in the most fruitful way* by attributing the [relevant] cognitive abilities" (Newen & Bartels, 2007, p. 294; emph. added). So success in certain cognitive tasks in which these distinctions matter is crucial.

The attractiveness of such comparatively modest accounts of concepts, wherein they are "representations posited to explain certain cognitive phenomena including recognition, naming, inference, and language understanding" (Piccinini, 2011; see also Piccinini & Scott, 2006), is exactly this: that they allow us to *explain* the behaviors of humans and other animals in a unified way.[8] The required level of abstraction and the connections to success in cognitive tasks allow us to distinguish between concept possession and 'blind' stimulus–responses, even when stimuli can be grouped into equivalence classes by the purported carrier. But then, *only* if there really *are* these cognitive phenomena to be explained, should we contemplate postulating concepts (see similarly Camp, 2009, p. 278).

Furthermore, if they are so to explain observed behaviors, we may associate a certain *stability* to concepts (cf. Camp, 2009, 277 ff.; Machery, 2009, pp. 23–24; Newen & Bartels, 2007, p. 294): It is the multiple applicability of the same concept THREAT by dog *D* that allows for the comparison between its behaviors towards *M* and *N* (see Camp, 2009, p. 279). On the other hand, concepts, unlike 'purely perceptual states' are also "revisable as a result of [...] a range of different experiences". This is one aspect that makes them distinctively *cognitive* (Camp, 2009, p. 279). Another is the

[8] As can be seen, I here presuppose a kind of realism about concepts, and hence bracket issues of meaning-skepticism, such as Wittgenstein's. Note, however, that it is in principle conceivable that at least some biological organisms should also rather be seen as having FCPs than concepts. While there is some case for concept-possession in insects (see, e.g., Camp, 2009), the case might be harder to argue for plants, and so the notion of an FCP could also be useful for describing their behaviors and activities.

fact that they are often conducive to the achievement of certain goals set forth by their carrier, which conduciveness they exhibit in virtue of their combinability with other representations of the same type (think dog-example again).

So concepts, modestly conceived, are relatively stable, revisable, and at least minimally semantic representations that explain certain cognitive phenomena and, particularly, the behavioral successes (or sometimes: failures) of humans and other animals. Now, given that machines programmed in terms of DNNs are apparently capable of succeeding (and sometimes: interestingly failing) in tasks such as image recognition or language processing, why hesitate to attribute concepts to them?

2.2 Consciousness and Semantic Plasticity: Pleas for Caution

My account of the specialness of DNNs (and other, sufficiently rich ML systems) in science will embrace the idea that we can associate conceptual meaning to them, but it will be just a bit more cautious than to simply claim that DNNs do in fact develop concepts. The reason is that, given the present state of the field, I am hesitant to fully embrace anthropomorphic notions in the context of AI (somewhat pace Buckner, 2018, 2021). Watson (2019) offers some ethical reasons for caution about such anthropomorphisms, but I believe there are also salient 'alethic' reasons for this. To see these in some detail, let me first dig a little deeper into the notion of a *representation* that underlies the notion of a concept.

As a zeroth step, I would like to dispel a distraction. For, there are two ways in which DNNs could be associated with representations: They could (a) themselves *be* scientific representations, much in the same sense as traditional scientific models; or they could (b) be said to *have* representations, much in the same ways as humans and other animals do. The first sense was recently disputed by Boge (2021, p. 51): We do not assign *meanings* to the formal elements of the function $\boldsymbol{f_\theta(x)}$ as we would do for the terms contained in some scientific model. Hence, while the elements of said model are used to represent properties of an oscillating system, the weights and biases contained in $\boldsymbol{f_\theta(x)}$ are not *used*, by researchers, to represent anything about the system of interest.

One might still uphold that the function $\boldsymbol{f_\theta(x)}$ *as a whole* is a representation of certain aspects of the system on which the data $\boldsymbol{x}$ were taken (e.g. Freiesleben et al. 2022, p. 9); and this is actually consistent with considerations found in Boge (2021, p. 55). However, this would mean establishing a rather limited sense of representationality; and for the present purposes, this sense is even irrelevant: We are, indeed, interested here in the question of whether DNNs can be said to have concepts, and even on a minimal account, this requires them to *have* representations.

Such representations *had* by a cognizing system are usually referred to as *mental* representations. A first, obvious reason for skepticism is hence that the relevant sense of 'representation' involves a notion of *mentality*, and that mentality is often assumed to bear some connections to *consciousness*. Now DNNs are, of course, *implemented* in (partly silicon-based) machines, and strictly *excluding* the possibility of consciousness emerging in such systems might be considered carbon chauvinism. But the point is not one of impossibility, but rather of there being little evidence that present-day AI

systems *actually are* conscious. I believe that most readers will agree with me, as evidenced by the discussion over Google's LaMDa mentioned above.

Actually, I submit that conscious content sometimes interacts with concept in such ways that the conscious content itself bears explanatory relevance for the kinds of cognitive phenomena that concepts are supposed to explain. For example, having a certain concept of TRIANGLE might enable me to draw certain inferences directly from visual introspection without being able to fully verbalize them: I might mentally vary the lengths of an imagined triangle's sides and immediately, from that introspective act, infer that angles must sum to a constant. Or I might visualize the Pythagorean theorem and thereby convince myself of its intuitive validity.

However, such a connection need not *always* be present: Several authors distinguish between explicit and implicit, or 'tacit', representations (e.g. Davies, 2015, for an overview). The latter ones are supposed to underlie certain apparently cognitively undergirded behaviors without necessarily entering into any specific relation to either linguistic verbalization or conscious content (Orlandi, 2020, p. 107).

Typically, tacit representations are assumed to reside on a *sub-personal* level (cf. Rescorla, 2020; Ryder, 2019). This is not really the same as detaching them from personhood altogether. Hence, insofar as personhood involves conscious experiences, one might still express reservations about entirely detaching tacit representations from consciousness. For instance, most authors seem to accept that for x to be a representation of y, x needs to 'be about' y (see Orlandi, 2020, for some amount of overview), and so representation may require *intentionality*. Furthermore, some (such as Kriegel, 2003; McGinn1988; Searle, 1992) have famously argued that even unconscious intentionality is ultimately *grounded* in consciousness, and so there is a reasonable stance that denies the possibility of mental representations without any consciousness at all.

But one may certainly refuse to accept such a connection to consciousness and the notion of tacit representation certainly allows the *possibility* of an a-personal, non-conscious entity with mental representations. So can we at least say that DNNs possess tacit concepts?

Brooks (1991, p. 149) gave a negative answer, based on the fact that whatever is there in AI lacks semantic content; something often imposed as a minimal requirement not only on concepts, but more generally on decidedly mental representations (see Ryder, 2019, p. 234). 'Semantic' can, of course, be fleshed in various different ways (say, as requiring reference, intensionality, etc.). But typically, it means at least a contentfulness that is associated with "conditions of satisfaction of some sort." (Hutto & Myin, 2020, p. 82) That is, mental representations "specify a way the world is such that the world might, in fact, not be that way." (ibid.)

Whether the verdict that AI systems cannot acquire semantic representations in this sense is still true today is of course a subtle issue: Interpretability methods, such as those discussed below, seem to suggest the presence of semantic content in modern DNNs. But the question remains whether the fact that *we*, human beings, can represent the goings on in a DNN in meaningful ways implies that they already 'have' meaning (see also Brooks, 1991, ibid.; Boge, 2021, p. 50).

That ex post interpretability in terms of mental representations and, specifically, concepts is not the same as concept-possession, was also already argued by Clark (1993). Clark claimed that what is needed for an AI system to possess concepts is the

ability to learn what he called *structure-transforming generalizations*: generalizations which "involve not just the application of the old knowledge to new cases but the systematic adaptation of the original problem-solving capacity to fit a new kind of case." (Clark, 1993, p. 73)

For instance, depending on the specifics of the training data, architecture, and even the loss function guiding the training, a DNN might fail to establish the dependency of DOG on LEG, FUR, EARS, and so forth. Thus, when the image to be analyzed constitutes a novel problem-situation that requires making use of this dependency, it could be incapable of drawing several inferences usually facilitated by DOG. I will discuss relevantly similar, suggestive examples below.

At this point, defenders of DNN-representationalism could still counter that the more upstream neurons in many-layered DNNs *can* often be shown, or at least reasonably assumed, to specialize to such less involved concepts (e.g. Goodfellow et al., 2016, p. 6; and below). However, in more complicated examples of a similar guise, to be discussed in Sect. 3.2), this is likely not correct. Much depends on whether the other concepts in question can be said to be *components* of the relevant object corresponding to the given concept in, say, an image (as in the DOG-example), or whether they are *semantically* constitutive of it in a more abstract sense.

Frankly, it is not just the dependency of concepts on simpler concepts, but rather their *connectibility* to other, similarly involved concepts that is crucial. Intel Labs vice president Gadi Singer illustrates the point as follows:

> A concept is not inherently bounded to a particular set of descriptors or values and can accrue almost unlimited dimensions [...]. For example, biology students signing up for their first class on epigenetics may know nothing about the field beyond vaguely recognizing that it sounds similar to "genetics." As time goes on, the once very sparse concept will become a lot more multifaceted as the students learn about prions, nucleosome positioning, effects of diabetes on macrophage behavior, antibiotics altering glutamate receptor activity, and so on. This example contrasts with deep learning, where a token or object has a fixed number of dimensions. (Singer, 2021)

'Dimensions' here, as above, mean independent features associable to the given concept. But some of these may characterize relations to other concept-like representations, as the example shows. Hence, an important element of the systematicity of concepts is their *plasticity* (which is one sense of revisability): A concept can be enriched by connecting it up with other, different concepts. Note that the plasticity in question is *semantic*: One can enrich a concept by connecting it up with other *already meaningful* representations—not with any old mental representation such as, say, spontaneous, random visual flashes before one's inner eye.

Now, following the above quote, any potential element of a DNN that could possibly realize a concept—such as a hidden unit, a hidden layer, or a pattern of activations distributed across units in multiple layers—would be severely limited in this respect by the number of connections it can possibly enter into (by the DNN's fixed architecture). But the same is probably true, to some extent, of the limitations imposed by biological

brains: The large and variable, but still finite, number of neurons and axons limits a biological organism's capacity for enriching its representations.

However, present day DNNs are restricted in a much more important way, namely by common training-procedures: Minimizing a certain pre-defined loss function means realizing *one* objective[9]; and this is arguably insufficient to accomplish the plasticity associated with concepts, which makes them so useful in navigating changing environments. Thus, unless there is a radical change in how DNNs are built and *trained*, it remains at best unclear whether they indeed establish the relevant relations that would justify concept-attribution, even when they *appear* to succeed based on conceptual reasoning.

I have thus sketched two independent lines of reasoning—one connected to consciousness, one to semantic systematicity—that suggest some skepticism towards the notion that even present-day DNNs have concepts, rather than just being *humanly interpretable* in terms of these. When I turn to concrete study cases below, I will put especially the second one to work. However, I clearly do not claim to definitively settle the matter here. All I am urging is *caution* with concept-attribution to AI systems.

Such a cautious approach allows for a rather unified view of DNNs as scientific tools: Large chunk of the technical literature certainly read as if we should take seriously the notion that DNNs are cognitive agents that develop internal models and representations of their environments. But a similarly large chunks read as if the present state of ML is nothing but clever, heuristic statistics. Allowing that DNNs can merely develop *proxies* for concepts—which could literally just be patterns of values the functions concatenated to give back $\boldsymbol{f_\theta(x)}$ take on—makes these views compatible: it is not overly demanding on the cognitive science-side but also not overly dismissive of DNNs' achievements.

Furthermore, I shall admit that with things like multimodal inputs on the horizon for systems like Google's PaLM,[10] we can envision a stage in the not-too-distant future actual concept-attribution to DNNs becomes a lot more defensible (see Clark, 1993, Chapter 4, for similar qualifications).

2.3 Functional Proxies

Despite all the skepticism, I also believe that *present-day* DNNs can develop *something* that plays the same role in classification, prediction, language processing and further 'cognitive' tasks as do concepts in human reasoning. Hence, how should we properly speak and think of this 'something'? In this section, I shall suggest a framework for this, by defining the notion of a '(functional) concept *proxy*'; one that can do all the work I expect it to do.

Consider what it means for some x to be a proxy for y. We usually do not mean by this that x can replace y tout court. Rather, we have in mind a set of *contexts* within

[9] Although this objective could contain several factors, realized by different, added-up terms, and so might be said to factor into multiple objectives (like 'minimize the mean squared error and maintain a small sum over all squared weights'). But this is inessential; this is very much unlike being able to train one's arithmetic skill by doing calculations first and then training one's musical skill by playing the guitar.

[10] See https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/ (checked 08/22).

which x can do whatever y does. For instance, a proxy variable in statistics is a variable that can be used to measure a latent trait or variable, because it strongly correlates with said variable. However, this usually neither means that it correlates *perfectly* with said variable (Carter, 2020, p. 174), nor that it satisfies all the *causal* roles the variable does (Pietsch, 2021, p. 158). So the scope of the proxy variable's use is limited across both measurements and purposes. Similarly, a proxy can refer to someone you designate to fill in for you in a decision-making process within a company or institution you are part of; but said person clearly doesn't thereby fulfill all the other roles you occupy in the institution.

Thus, as a first (working) definition, I will say that, given a set of contexts, C, then x is a *proxy* for y, relative to C, *iff* x occupies the same roles as does y in all $c \in C$, but does not do so in some $c' \notin C$.

Now, 'roles' can mean lots of things: In the decision-making case, they are legal roles, in the statistical case, they are inferential roles. What roles could proxies for concepts play? I submit that the relevant roles are *causal* ones: Entertaining a certain concept of DOG might stimulate me to say 'look, that cute dog over there', whenever certain constituent stimuli are present. Similarly, it will stimulate me to infer that I can likely steer the behavior of the object constituted by these stimuli by exclaiming things such as 'sit' or 'roll'. Hence, the presence of the concept DOG causally contributes to my observable behavior and to the 'outputs' I produce; though indirectly via the inferences and other cognitive achievements to which it contributes and which result in said outputs.[11]

Causal roles are usually identified as that which determines the *function* of something (see Levin, 2018, §1). Hence, I shall call a proxy x that fulfills all the same causal roles as some y in all the c in a set of contexts C a *functional* proxy for y.[12]

Now, as for the definition of a *concept* proxy, the contexts C that matter may be characterized as reasoning, or, more generally: cognitive, *tasks*, T. These may comprise classification, i.e., sorting encountered entities under pre-defined classes, categorization, i.e., finding new classes for these,[13] inferring inductively into the future or to a generality, and so on. Putting these ideas together with the minimal account of concepts appealed to above, I define FCPs thus:

> Given a set of tasks, T. Then x is a *functional concept proxy (FCP)*, relative to T, *iff* in any $t \in T$, but not in some $t' \notin T$, x fulfills all the same causal roles as does any relatively stable, revisable mental representation y posited to explain certain cognitive phenomena and behavioral successes exhibited by its carrier.

This is a deliberately permissive definition, as it should be, given that FCPs are supposed to be something that can be had by what is, under a slightly dismissive description, 'just' a parametric function. However, an anonymous referee has confronted me with the following set of interesting questions:

[11] See Piccinini (2022, p. 5), for similar views.

[12] The importance of the causal roles shall also become clearer in Sect. 2.4 below.

[13] For an example, see (Knüsel & Baumberger, 2020) I take 'categories' to be basic classes in classification systems.

1. Would a linear regression model with a term for "socio-economic status" possess an FCP for SOCIO- ECONOMIC STATUS? If not, why not?
2. Would an automatic door equipped with an electric eye possess an FCP for person? If not, why not? And what if it was trained with Reinforcement Learning?
3. Would a GOFAI program like Winograd's SHRDLU possess an FCP for block or pyramid? If not, why not?
4. Would a discriminative method like a support vector machine that classifies dogs from non-dogs possess an FCP for dog? If not, why not?
5. Does possession of FCPs require some kind of "interpretable" substructure like features in hidden layers, or could a discriminative method with a complex decision boundary possess FCPs?

I would answer 1. as follows: "Presumably yes, but because we have put it in by hand". The distinguishing characteristic is that DNNs are, according to the evidence discussed below, capable of *developing* FCPs themselves. I fail to see how this could be possible for the regression model, as the term was assumed to have been *handcrafted* to represent the socio-economic status of people.

To 2., I would respond: "Presumably yes, but again because we have crafted it in this way." Furthermore, using Reinforcement Learning, it might even be *conceivable* that the door develops FCPs for *things we had not designed it to recognize*. But whether this is *plausible* depends on whether we can gather positive evidence to this effect—as is possible with DNNs.

3. Is a bit more involved, so I will return to it below.

To 4., I would respond: "Since a support vector machine is a kind of (shallow) neural network (Baldi, 2021, 13, pp. 56–57), it is certainly thinkable (given sufficient length) that it develops FCPs, and even for things it was not explicitly trained to classify (say, ears and tails)." So, again, FCPs and even their development may not strictly be restricted to *D*NNs. However, whether support vector machines do or do not develop FCPs depends on whether we can relate, say, the values taken on by a non-linear kernel (the machine's activation function) to meaningful elements in an image (see below). And to my knowledge, we happen to have positive evidence for this only in the case of DNNs.

This brings me immediately to 5., to which I respond: "While this is a typical way of identifying FCPs (see below), this may not be necessary." As the definition says, there just needs to be 'something' that fulfills the same roles as a concept. Thus, whether we can attribute FCPs or not depends on whether we can gather evidence that a given system can exploit information in a certain way—and this does not necessarily require that we can identify that something in terms of some interpretable structure. It only requires, much in the same ways as this is the case with actual concepts, that we have reason to postulate the FCP's existence, given the system's performance.

I believe that this definition is also sufficient for distinguishing FCPs from actual concepts. For example, allowing a more involved definition of 'concept' for the moment, which requires at least some grounding in consciousness, a given task may involve *imagining* an object and drawing inferences based on the given mental image. So unless DNNs become conscious, this would be impossible for them. However, even disregarding these more involved issues, a general pattern for identifying mere

proxyhood emerges: It might be possible, by a slight alteration of the given task, to show that the FCPs attributable to DNNs are quite likely missing relevant links to other concept-like representations—and even feature links to stuff that quite clearly lacks meaning.

These missing or erroneous links can be exhibited by means of the mistakes prompted by actual or merely contemplated alterations of the tasks DNNs are subjected to, as I shall argue below. This means showing that the relevant sort of semantic systematicity and plasticity typical of concepts is missing, by taking a DNN out of its comfort zone (the $t \in T$). However, a more direct route might be possible as well, which consists in straightforwardly showing the questionability of the meaningfulness of activation patterns by taking a DNN out of its comfort zone. I here have in mind the notorious problem of adversarial examples, addressed in Sect. 3.3.

Finally, note that there could also be an extended set of tasks $T^* \supset T$ and tasks $t^* \in T^* \setminus T$ in which a DNN succeeds by means of its FCPs, but no human being does, using only the concepts she has available.[14] That is clearly permitted by my definition and it hints at the main problem for science I am embarking upon here: That DNNs and other, similarly complex ML systems may selectively outsmart us despite not (yet) having actual concepts, and that this may put us at a loss when it comes to an understanding of the subject matter.

2.4 Evidence for the Existence of Self-Developed FCPs

Why think there is such a thing as self-developed FCPs (if not concepts) in DNNs? The fact of the matter is that there is some amount of empirical evidence for this, although the distinctions I have drawn above have of course not yet been acknowledged in the relevant literature.

Before going into relevant studies, note that there is also *textual* evidence for the relevance of concepts for understanding Deep Learning successes. For instance, the very notion of *representation learning* builds around this: It is generally assumed that DNNs' hidden layers are capable of learning distributed representations (Goodfellow, 2016, pp. 536–537), where these representations are indeed typically understood in terms of concepts:

> When we speak of a distributed representation, we mean one in which the units represent small, feature-like entities. In this case it is the pattern as a whole that is the *meaningful* level of analysis. This should be contrasted to a *one-unit-one-concept* representational system in which single units represent entire *concepts* or other large meaningful entities. (Hinton, 1986, p. 47; first and third emphasis mine, second original)

Thus, the idea is that, through the iterative updating of its parameters, a DNN can acquire a certain concept if its hidden layers learn to specialize to representing certain

[14] Here, $t' \notin T^*$, i.e., the tasks on which the DNN fails correspond to a yet distinct set, T'. So in other words, the scope of tasks on which FCPs *and* concepts both yield successes correspond to an overlap between two distinct sets ($T' \cap T$). I owe thanks to an anonymous referee for pointing out that this is a nice way of clarifying FCPs and, frankly, their relation to the ASHP. Note also that this makes FCPs simply *different* from concepts, not necessarily derivative of them, nor generally inferior.

features, so that the overall pattern of activation may signify these features' presence or absence, respectively.

In the philosophical literature, Cameron (Buckner, 2018, p. 3) has recently similarly suggested that DNNs are capable of forming "subjective category representations or 'conceptualizations'", through a process he calls "transformational abstraction". Likewise, López-Rubio (2020, p. 3) argues that "emergent visual concepts are learned spontaneously by [...] deep networks because they are useful as intermediate steps towards the resolution of the final goal [...]." Overall, there appears to be a broad consensus, both in technical and relevant philosophical literature, that DNNs are capable of forming something akin to concepts. Understanding the *limitations* in attributing actual concepts to DNNs, however, requires looking carefully into the details of some *non-textual* evidence.

Consider first the study by Bau et al. (2017), also discussed by López-Rubio (2020). In this study, Bau et al. (2017) introduced a method they called 'network dissection', which aimed at mapping out the extent to which activations of individual hidden units of several convolutional DNNs align with humanly interpretable concepts at multiple scales, such as color concepts, object concepts, scene concepts, and so forth. To this end, a dataset with a broad range of images of different scenes or objects was used, wherein each image is attached with various labels down to the pixel level (specifying the color, but also the object to which the pixel belongs). These images were also equipped with annotation masks, which can be visualized as a dimming of every pixel that does not belong to a given object falling under some concept.

To quantify how much individual hidden units would align with this humanly interpretable segmentation, Bau et al. (2017) defined a binary activation map, based on hidden units' activations that were so high that they were exceeded in only half a percent of the images by the given unit. The interpretability of some unit in terms of a given concept was then evaluated with the aid of the intersection-over-union measure over all images, which basically computes a 'matching-percentage'.

Bau et al. (2017) then defined those units as interpretable for which a set of independent human raters agreed with the 'ground truth' in a yes/no decision, i.e., with the labels as given by some annotation mask. These ground truth labels were also checked for consistency by asking a second set of raters. Both the agreement between human raters about the ground truth labels and the agreement between the activation mask and the concept was the highest for later convolutional layers, which are typically specialized to object rather than color or edge recognition. An exemplary illustration is provided in Fig. 2

A second study by Bau et al. (2018) probed even deeper into DNNs' conceptual interpretability, which is suitable also for highlighting several reasons for considering such interpretable activations (or patterns thereof) *functional proxies* for concepts, rather than actual concepts. In this second study, the network investigated was the generative part of a 'GAN'; a generative-adversarial network. In a GAN, there is a generative part, G, that is trained to produce images (or other data-like outputs) and an 'adversarial' part that tries to decipher whether a given instance $\boldsymbol{y}$ is G's output or a genuine data instance (say, a photo taken with a camera). This type of architecture can be used either to produce ever better 'fakes', or to identify fabricated data such as machine-generated images (Goodfellow et al., 2014a).

Fig. 2 Exemplary units of the ResNet image recognition DNN interpretable in terms of the concepts HOUSE and DOG respectively. Adapted from Bau et al. (2017) under a CC BY 4.0 license. Colour available online

By 'dissecting' the generative part of a GAN in partly the same ways as with the image-recognition DNNs discussed above,[15] Bau et al. (2018) could not only demonstrate a match between activations of hidden units and certain concepts, but also the *causal* relevance of those units for the presence of conceptually meaningful image patches.

In particular, Bauet al. (2018, p. 5) used a set of *interventions* on hidden units—that is, precise, selective manipulations of their values—to test the effects of changes to these units on the generative DNN's output. Interventions on certain variables (such as a hidden unit's activation) are often held to be key to elucidating their causal relations to other variables (see Woodward, 2003)—such as a generative DNN's output. Thus, the results of Bau et al. (2018) can be used to substantiate the *functional* aspect invoked above: Recall that the whole point of inserting the qualifier 'functional' into my definitions in Sect. 2 was to elucidate *what roles* an x present in some DNN needs to play in order to qualify as a proxy for some concept, relative to the tasks we subject the DNN to.

In particular, after identifying (sets of) conceptually meaningful units, Bau et al. (2018) could show that ablating these units, i.e., setting their values to zero by hand, removed the corresponding parts in the generated images. For instance, ablating more and more units that had been identified with TREE, the generative DNN could be shown to produce images with less and less trees.

Now, as matter of fact, studies on human beings

> that have probed for knowledge of particular concepts across different modes of access and output (e.g., fluency, confrontation naming, sorting, word-to-picture matching, and definition generation) demonstrate that patients with Alzheimer's disease are significantly impaired across all tasks, and there is item-to-item correspondence so that when a particular stimulus item is missed (or correctly identified) in one task, it is likely to be missed (or correctly identified) in other tasks that access the same information in a different way [...]. (Salmon, 2012, p. 1226)

[15] Some changes are briefly crossed below.

Hence, the kind of brain-damage related to Alzheimer's disease apparently leads to the loss of certain concepts in human beings. By the same token, an imagined future 'evil neuroscientist' might selectively inhibit neural activity in a biological brain in precisely such ways that a measurable loss in relevant cognitive abilities would result, suggesting that the relevant concept had been 'deactivated'—in analogy to the study by Bau et al. (2018). This justifies the relevance of the *causal* roles played by, viz., the functioning of, concepts in the reasoning and cognition of biological organisms.

It doesn't at all clear up the requirement for merely speaking about *proxies* though. First note that the fact that these hidden units' activations can so fulfill a relevant causal role in generating images of, e.g., trees *when the DNN is supposed to* generate trees, is evidence enough for *candidate* proxyhood: The activation patterns *can* fulfill the same causal roles as do concepts in tasks wherein the respective carrier is supposed to create a visual representation[16] of a tree (in the relevant contexts, C). However, in order to show that they are *just* proxies, it is necessary to show that they do not fulfill said roles across *all* tasks wherein a given concept would.

Recall that I had claimed both an element of stability as well as of plasticity to concepts' systematicity: They remain stable enough so that the same concept may be said to combine with other concepts over different instances in time, and are plastic enough so that the given concept can be enriched by being equipped with further connections to other concepts.

In order to realize specifically the plasticity aspect of this, it would be necessary to enrich the activation patterns aligned with concepts in these two studies by connections to (or co-activations with) further hidden units, so that the representational capacity of the DNN would be increased. But besides the aforementioned general limitations to this imposed by architectural and learning-related constraints, I shall here provide some reasons for thinking that such co-activations and connections can actually be shown to lack the relevant *semantic* features.

The second study by Bau et al. (2018) can, in fact, be used to advance just such a reason: In addition to the causal investigation of units' contributions to humanly interpretable pixel-patterns, Bau et al. (2018, p. 2; emph. added) noted that their "method can diagnose and improve GANs by identifying *artifact*-causing units". In particular, correlating certain units with artifacts in generated images and then ablating these units contributes "to debugging and improving GANs" (ibid.).

However, consider the type of artifact typically in need of such 'debugging': Typical artifacts recognized by Bau et al. were patterns of vertical bars or smudges of greyish-violet color. Furthermore, the improvement of the GAN proceeded not by 'educating' the generative part further, but by *ablating* those artifact-causing units (cf. Bau et al., 2018, p. 13).

Now, it is not typical, say, for a bed to co-occur with either a set of vertical bars or greyish-violet smudges; hence, a human being would likely never learn a semantic connection between BED and GREYISH- VIOLET SMUDGE, in conceiving of the interior of a bedroom. However, greyish-violet smudges—overlayed with other patterns, so as to be invisible to the human eye—might be typical co-occurrents with *sets of pixels*

[16] Note that in a human being, this could also just be a mental image: Having a certain concept of trees will certainly prompt me to imagine a tree when an instructor mentions the relevant word.

in RGB images that, to a human being, represent beds. But if greyish-violet smudges appear as meaningful to the DNN as beds, it becomes doubtful whether anything is meaningful for it at all. Hence, there is reason to think that the generator part of the GAN had really only learned a statistical correlation between *sets of pixels*, instead of acquiring a concept of beds.

To make this just a bit more plausible, recall how the success of parrot Alex was *best explained*, according to Newen and Bartels (2007), by attributing a minimal semantic net to Alex. However, when beds are connected to greyish-violet smudges as elements typical of bedrooms by a generative DNN, it becomes unclear that concept-possession is indeed *the most fruitful way* to explain its generative abilities: Any purported semantic net would then have to include connections between bed-like objects and meaningless blobs. It would thus be anchored in something which is *not* an object, though aligned along the object-dimension (or: located in the object-subspace) of said net. This doesn't sound very convincing.

I submit that it seems much more plausible to assume that the DNN merely learns to exploit correlations between pixel-patterns, and that there *is no* semantic net present within it: The patterns learned by DNNs are not *contentful* representations, and their potential 'satisfaction conditions' are really exhausted by the respective optimization method terminating near some minimum of a loss function. Like the Google engineer who arguably fell prey to the *illusion* of sentience, we thus arguably fall prey to an illusion of there being meaningful representations attached to DNNs, when they skillfully learn to exploit (and reproduce) statistical patterns. This illusion is exposed, however, when we pay careful attention to the kinds of mistakes DNNs make.

It is easy to see that this evidence against actual concepts in DNNs correlates with the kind of task, T, we subject them to. For instance, imagine that the generative DNN had been tested on a range of commands that had simply happened not to stimulate the smudge-producing hidden units. Then it would have reproduced bedrooms just fine, and success would have been granted. Similarly, consider a set of tasks, T, in which a generative DNN trained in the ways discussed above was supposed to produce images that fool human beings into believing they are real camera footage. Clearly, here the generator could easily fail, because the regular appearances of greyish-violet smudges might raise suspicions in suitably educated test subjects.[17]

I acknowledge that defendants of the attributability of outright concepts to DNNs could maintain that the DNN has, among other things, learned the object-level concept GREYISH- VIOLET SMUDGE. But I claim that this would mean stretching the notion 'concept' too far: Extant theories of concepts individuate them by their meanings as well as the connections to other already meaningful representations, and a major reason for postulating concepts is explanatory power. It seems rather contrived to associate meanings to hidden units producing greyish smudges when the respective DNN's behavior can equally well be explained by learned statistical correlations between pixel-patterns. I consider the foregoing to deliver a sensible amount of justification for preferring to speak of mere functional proxies for concepts, and will return to the matter in Sect. 3.3.

[17] See Marcus and Davis (2020) for a discussion of similar mistakes in natural language processing DNNs.

3 Success and the Novelty of DNNs in Science

3.1 DNNs Versus Traditional Multivariate Analysis and 'GOFAI'

So far, I have only made a case for the existence of self-developed FCPs (and against actual concepts), but neither for the fact that they enable success nor that they make DNNs special. Let me begin by first looking into the general connection between DNNs (or ML more generally) and statistics a little more carefully. That there is some kind of connection probably goes without saying (see Flach, 2012, xv; Goodfellow et al., 2016, p. 95; Skansi, 2018, v).

There is, however, some disagreement about the exact connection between statistics and ML, not least when it comes to fundamental matters. For instance, Boge (2021) has recently argued that both statistical models and DNNs are in a sense not explanatory, whereas Srećković et al. (2021) argue that statistical models are more explanatory. The apparent disagreement can be resolved, however, when one looks into the details.

In essence, statistics might be characterized as an activity of collecting data samples $\vec{\boldsymbol{x}} = \langle \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \rangle$ and, more often than not, using them to infer something general about a 'population' from which the data were drawn, or something about future samples. Usually, this is done with the aid of parameterized (probability) models $P_{\boldsymbol{\theta}}(\boldsymbol{x})$ that, for some choice of $\boldsymbol{\theta}$, match the data's frequency distribution, or the frequency distribution of a function of $\vec{\boldsymbol{x}}$ (a 'test statistic'), in the sense specified by an appropriate criterion for the matching. However, the details of this process can vary drastically.

A major conceptual difference has been recognized by various statisticians from at least Neyman (1939) between *theory*- and *data-driven* approaches to statistics. As Neyman (1939, p. 55) writes, applying statistical concepts to data requires "some system of conceptions and hypotheses, the consequences of which are approximately similar to the observable facts." However, "this similarity may be differently placed"; it could either apply

> to the shape of [relevant probabilistic] curves and to the shape of the empirical histograms. Otherwise it may apply to certain real features of the phenomena studied and to some mathematically described model of the same phenomena. And if the theoretical distributions deduced from the mathematical model do agree with those that we observe, and if that agreement is more or less permanent, we say that the mathematical model has "explained" the origin of the distributions. (ibid.)

Thus, whenever one has a theory or theoretical model in hand, said theory or model may be used to determine the *expected* empirical distribution of data, and statistics can serve the aim of testing the theory. The use of statistics may here either reflect the theory's stochastic nature, or the noisiness of the measurement conditions, or both (Lehmann, 1990, p. 166). In turn, if the theory thus reasonably matches the data, it may be said to explain the observed phenomena.

Data-driven approaches to statistics, in contrast, usually serve the goal of 'mere' prediction:

> For example, in trying to predict whether a customer will buy a particular item next week, one does not base one's prediction on a set of differential equations [...], but rather on a (probably fairly simple) [...] empirical model [...] relating past purchases to the characteristics of the customers making them. (Hand, Hand (2009), p. 294)

Such a fundamental distinction between uses of statistics has, in some form, been acknowledged and echoed by several statisticians (see, for instance Breiman, 2001; Davies, 2014; Hand, 2009, 2019; Lehmann, 1990; Shmueli & Koppius, 2011,). Furthermore, we can see that ML has a lot in common with, or is even an instance of, *data-driven* statistics[18]; and the disagreement between philosophers such as Srećković et al. (2021) and Boge (2021) is resolved when one realizes that the former focus on theory-driven uses whereas the latter is focused on data-driven ones.

Of course, there are also differences in detail between ML and traditional data-driven methods in statistics, starting with the fact that the output of a DNN $\boldsymbol{f_\theta(x)}$ need not (though it might) be a probability distribution: In the case of AlphaFold2, it is a rotatable depiction of a three-dimensional protein shape. Nevertheless, the generation of the final $\boldsymbol{f_\theta(x)}$ is generally a statistical *procedure*, as it involves adapting a certain function by using probabilistic methods and random samples of data. Furthermore, typical choices of activation functions in downstream layers have probabilistic interpretations (Goodfellow et al., 2016, 178 ff.), and so the output of a DNN may quite generally be considered the *most probable* class label (or: protein-shape, reconstruction of the data,...), on account of a statistical model 'hidden' within the DNN.

Finally, the learning process is fundamentally described in an entirely statistical vocabulary (e.g. Shalev-Shwartz & Ben-David, 2014). But very often, methods are used without regard to some formal justification. For instance, the shape of many regularizers is not directly motivated by traditional statistics, and their effects are often understood only to a limited extent (e.g. Moradi et al., 2020).

In sum, there are several differences between ML and (data-driven) statistics, in the general style of models, what they can achieve, and how one treats them. I maintain that all these differences are not really fundamental though: Both DNNs and traditional data-driven statistical modeling are methods for analyzing data and inferring something from that analysis, and all ML models at some point appeal to techniques that were chiefly developed within statistics. Yet, a core fundamental difference lies, I believe, exactly in the presence or absence of self-developed FCPs.

However, DNNs are not just used as analysis methods, but considered instances of AI. Hence, might a better pick for comparison not be 'Good Old-Fashioned AI' (GOFAI) systems, where "GOFAI methodology employs programmed instructions operating on formal symbolic representations", and "A GOFAI symbol is an item in a formal language (a programming language)" (Boden, 2014, p. 89)? This brings me back to the reviewer question 3., mentioned in Sect. 2.3.

[18] This is not so clearly true anymore in integrated approaches that also use theoretical information next to data, as suggested, e.g., by Reichstein et al. (2019) for earth science and true to some extent also of AlphaFold2; but apart from the considerations on inductive bias offered in Sect. 1, I will bracket this issue here.

Note, first, that I am interested here in whether DNNs are somehow special within scientific applications. From that vantage point, of course the technological advancements brought about by the GOFAI-approach must be acknowledged, which have certainly impacted science in many direct and indirect ways (see Boden, 2014, p. 101). However, is an analogous problem to the ASHP posed by GOFAI, wherein computers leap ahead of us in such a way that they can accurately predict, classify, and discover, while human researchers have a hard time understanding the predicted, classified, or discovered?

I believe this is doubtful, at the very least due to questions of extent, and I will illustrate this using examples discussed by Dreyfus (1992, including the one suggested by the reviewer). First, consider the most plausible GOFAI candidate in Dreyfus's discussion for an AI system that could potentially develop FCPs: Winston's 'concept learning' program (see Dreyfus, 1992, 21ff.). "Given a set of positive and negative instances", Winston's program was able to, "for example, use a descriptive repertoire to construct a formal description of the class of arches." (ibid.) Furthermore, since said program was not crafted ab initio with some sort of representational means for representing arches, it may be claimed to have developed an FCP, relative to the task just described (offering formal descriptions). However, even if we accept this as true, there are reasons to be suspicious of the nature and scope of this FCP-developing ability:

> [...] Winston's program works only if the "student" is saved the trouble of what Charles Sanders Peirce called abduction, by being "told" a set of context-free features and relations—in this case a list of possible spacial relationships of blocks such as "left-of," "standing," "above," and "supported by"—from which to build up a description of an arch. (ibid.)

Thus, on the one hand, we might argue that the concept (or the FCP, frankly) was crafted in after all; even if only implicitly. That is, we might hold that it is only meaningful to speak of the (somewhat autonomous) 'development' of an FCP if this requires being able to react in fairly novel ways to a given problem set, and so *other than* by "put[ting] together available descriptions in such a way as to match these encountered cases". This latter sort of task could be seen as making explicit that the system in question *already had* a given FCP, *by design*.

If, on the other hand, we reject this line of reasoning, there would certainly still be a major difference in extent between what GOFAI systems were able to do and what modern DNNs are capable of, in terms of self-developed FCPs. Indeed, it is out of the question, for the very reasons given by Dreyfus, that Winston's system could have developed FCPs for objects not at all connected to the resources (descriptions) made available by the programmers. It seems that this is different in DNNs, as shown by the evidence given above and below. But maybe an advocate of the in-principle equality of GOFAI to DNNs on the grounds on which I am evaluating both here could at least argue that we need to alter the ASHP by including a qualifier such as 'in novel ways and to an unprecedented extent'.

Let us consider the reviewer's favored example, SHRDLU, now to see whether this verdict can be upheld. SHRDLU

> simulates a robot arm which can move a set of variously shaped blocks and allows a person to engage in a dialogue with the computer, asking questions, making statements, issuing commands, about this simple world of movable blocks. (Dreyfus, 1992, p. 5)

In the course of handling blocks and responding, SHRDLU could successfully disambiguate pronouns such as 'it', when multiple referents were conceivable, and use a deductive system to find an actual example for answering modally qualified questions (such as "can a pyramid be supported by a block?"; Dreyfus, 1992, p. 7). Does this ability to handle blocks and respond to queries by a user not speak in favor of FCP-possession on the side of SHRDLU? Maybe so, but as the discussion in Sect. 2.3 should have made clear, this by itself is not interestingly distinguishing (given the liberality of 'FCP'). Thus, does the fact that the deductive system can apparently alter SHRDLU's conception of 'pyramid' not imply the development of FCPs? I doubt it; at least for the 'ab initio' development that seems to be possible for modern DNNs. Furthermore, even if this was answered in the affirmative, I believe the difference in extent mentioned above for Winston's program could be even more so upheld in this case—thus making neither system interesting for the question of whether AI may be said to have a profound impact on science in the sense promoted by the ASHP.

3.2 The Role of FCPs in Generating Success

To make the case more clearly, let us first turn to the question of success now. In order to make a case for a connection between success and (the development of) FCPs, I may partly rely on authority again: Buckner (2018, p. 4) too argues that convolutional DNNs "are so successful across so many different domains because they model a distinctive kind of abstraction from experience", which, as we have seen above, he takes to result in 'conceptualizations'.

I will here not engage with the question of whether Buckner's account, which he takes to vindicate some empiricist themes in the philosophy of mind, is ultimately correct. This is a thorny subject and I cannot judge whether the process is not, say, better phrased in terms of Kantian 'spontaneity' (Fazelpour & Thompson, 2015), with its decidedly rationalist elements, or whether these views are ultimately even compatible (cf. Buckner, 2018, p. 12). Instead, with an eye on the discussion to follow, I will look into studies that provide evidence for the connection between success and the discovery or formation of 'higher level' concepts.

Some striking such evidence comes from particle physics, a field in which the analysis of massive amounts of data from particle colliders stimulates various new ideas in ML. It has here been recognized for a while that DNNs appear to be able to infer what physicists call 'higher level features' (Baldi et al., 2014; Chang et al., 2018). These are features that are determined as typically non-linear functions from other, 'lower level' features that are more directly read off from the data (Baldi et al., 2014, p. 3).

A typical example of a low level feature is the transverse momentum; the momentum-component a particle has transverse to the particle beam in a collider. This can be inferred from energy deposits particles leave in the detector, referred to

as 'raw data'. For instance, for a charged particle, 'particle trackers' apply a magnetic field and calculate the transverse momentum from the field strength, the radius of the particle's curved track in the detector and its charge, according to the Lorenz force law (see Albertsson et al., 2018, p. 7).

Given that physics information is needed to perform such 'reconstructions', this makes low level features "still high-level relative to the raw data" (Albertsson et al., 2018, p. 8). Nevertheless, for efficient event classification and analysis, physicists often rely on quantities that are still 'higher level', i.e., defined by complex inferential chains that rely on further physical principles. They are thought to "capture physical insights about the data" (Baldi et al., 2014, p. 2).

An example of such a higher lever feature is the reconstructed invariant mass of a decayed particle. Most particles produced in high-energetic scatterings will decay into more stable ones; for elementary particles, in the particle annihilation and creation processes predicted by our current quantum field theories. Then, given the relativistic energy–momentum relation and the conservation of energy and momentum, it is possible to reconstruct the mass of a decayed particle from the energies and momenta of measured particles it decays into.

A study that provided evidence that DNNs are able to *autonomously* infer the information contained in such higher-level variables was presented by Baldi et al. (2014). In this study, several hypothetical physics processes were simulated and the simulated data were then processed by a DNN. One such process included a more massive, electrically neutral Higgs boson, H^0, which decays into the known 'light' Higgs boson, h^0, that was discovered in 2012 (cf. Aad, 2012: Chatrchyan, 2012), via further, positively or negatively electrically-charged Higgs bosons, $H^{\pm}$. The DNN was now trained to classify events that contained the H^0 as 'signal' and events that did not as 'background'.

The DNN was now trained for this task using lower-level data such as the transverse momentum described above. Actually, higher level variables, such as the reconstructed invariant mass of decayed, intermediate particles, can expose the differences between background and signal data much more clearly, and so have higher discrimination power (cf. Baldi et al., 2014, pp. 4–5). Remarkably, however, feeding the higher-level variables to the DNN in addition to lower-level ones during training resulted only in a modest increase in performance, while training the DNN solely on the higher level variables actually led to a more drastic decrease as compared to when it was trained solely on lower-level ones. This behavior was in marked contrast to other classifiers used in the study, such as a boosted decision tree and a neural network with only one hidden layer (cf. Baldi et al., 2014, p. 7).

These results suggest that the DNN somehow autonomously discovers the information contained in higher-level variables. However, in a different benchmark with simulations including supersymmetric particles, the differences between the DNN and other classifiers were not as prominent (cf. Baldi et al., 2014, p. 8). Furthermore, the fact that a DNN trained only on higher-level features performs worse than with the lower level ones does not make the higher level variables' connection to success all that clear.

In this last respect, another study by Chang et al. (2018) is instructive, which was in many ways similar to that by Baldi et al. (2014) but added further ideas. Herein,

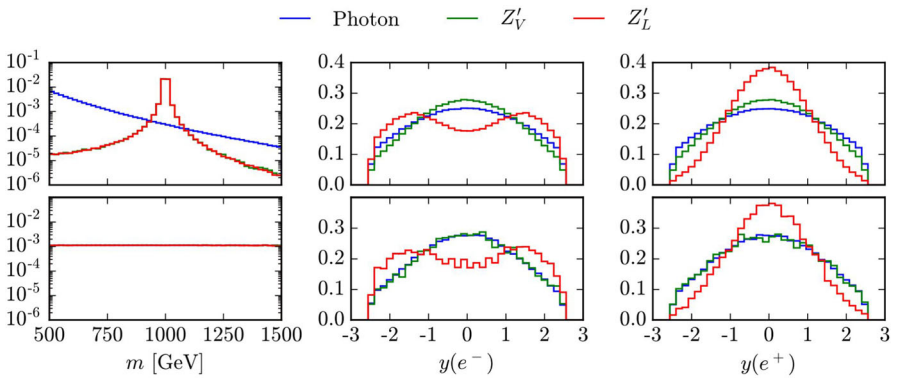

Fig. 3 Data planing illustrated on the example of a reconstructed particle mass. Reproduced from Chang et al. (2018) under a CC BY 4.0 license. Colour available online

the DNN was 'robbed' of the information on higher level variables after training, and corresponding drops in performance were observed. In an ingenious procedure called 'data planing', Chang et al. (2018, p. 2) removed the information contained in certain variables, effectively by weighting any given input variable $\boldsymbol{x}_i$, characterizing a certain scattering event i, by the inverse height of the histogram for the given higher-level variable at i.

This is illustrated in Fig. 3, for the reconstructed mass of a decayed particle: The upper panel shows unplaned histograms, the lower one planed ones. As is easily seen, the mass histogram itself (which originally has the characteristic 'bump' indicating a particle) is flattened out into a uniform distribution. But changes in other higher-level variables, such as the rapidity y for electrons (e^+) and positrons (e^-), are far more subtle, as shown in the mid and right plots.

The most important observation of Chang et al. (2018, p. 4) was that the performance of their DNN dropped significantly in response to the planing. To show this, Chang et al. (2018, p. 3) used two physics models to generate simulated data on which the performance was tested. In both models a new particle, called Z', was included, but only in one of them was it coupled with unequal strengths to known particles and anti-particles, such as electrons and positrons. In the case of the symmetric coupling, planing for the invariant mass of the Z' was sufficient to reduce the DNN's perfomance to guesswork. In the asymmetric case, another higher-level variable had to be introduced in addition, in order to achieve the same effect. This was the so-called rapidity difference, which provides information on the different angles into which electrons and positrons scatter, relative to the direction of the beam of colliding particles. In the case of uneven coupling, a difference in these rapidities is to be expected, and the network's performance indeed wound up no better than guesswork when the rapidity difference was planed away, whereas planing only for the mass left it in a still somewhat better place.

This study is impressive, as it quite clearly shows the dependency of the DNN's success on the presence of information on higher-level variables *in the data*. And as in the study by Baldi et al. (2014), the DNN may be said to have 'discovered'

this information 'on its own'. Furthermore, the variables planed for clearly encode physical concepts: that of a particle's mass, or that of the 'tilt' of its trajectory relative to a certain direction of reference.

Nevertheless, given everything said in Sects. 2.2 and 2.4, I believe it is utterly implausible to say that the relevant DNN had *literally* developed these concepts. For quite certainly, it had no conception of particles scattering and decaying at all, thus missing relevant semantic links to PARTICLE, SCATTER, and DECAY. Hence, had the relevant DNN been taken out of the comfort zone of the kind of classification task it had been trained for, and into one in which these concepts and their connections would have mattered, it would have clearly failed. More importantly, it would then probably have exhibited artifacts (or reacted to artificial features) similar to the ones discussed for the generative network studied by Bau et al. (2018) in Sect. 2.4.

To make this a little more plausible, consider once more Fig. 3. It is noticeable that there are, of course, also swift changes in the histograms for quantities *other* than the quantity planed for. The same extends to the lower level quantities that make up the entries x_{i_j} of data vector $\boldsymbol{x}_i$ for event i: If the frequency of vectors $\boldsymbol{x}_i$ with a certain, specific set of properties (such as the transverse momentum falling into a certain bin) is changed, then plotting events with said properties in a histogram will lead to a different result.

This makes it entirely reasonable to suppose that the DNN had here learned to specialize to these swift changes in event-frequencies in a highly effective manner, and it is also reasonable to suppose that there are activation patterns that correlate with humanly meaningful representations of these changes. In fact, this is not just reasonable, but in yet another study by Iten et al. (2020a), that too exhibited some relevant structural similarities to the studies discussed do far, this could be evidenced directly.

In said study, a specific encoder-decoder architecture, called SciNet, was used to "investigate whether neural networks can be used to discover physical concepts from experimental data." (ibid., 1) The precise setup used by Iten et al. is an instance of a generic DNN architecture called a *(variational) autoencoder* (cf. Iten et al., 2020b, for a brief overview), which compresses the data and then decompresses them again, where the intermediate, compressed layers are interpreted as developing a 'latent representation', and the output then identifies the network's 'interpretation' of the data based on this latent representation.

Surprisingly, when SciNet was used to predict, e.g., the future position of a pendulum from its past positions, it had learned "to extract the two relevant physical parameters from (simulated) time series data for the x-coordinate of the pendulum and to store them in the latent representation [...] [*w*]*ithout being given any physical concepts*"(Iten et al., 2020b, p. 16; emph. added): Out of three latent units in the most compressed layer, one unit's activation correlated perfectly with the damping-constant of the harmonic oscillator equation and another one with the spring constant, while the third unit was barely activated, meaning that it was superfluous (see Fig. 4). Hence, it seems that "SciNet has recovered the same time-independent parameters [...] that are used by physicists." (Iten et al., 2020b, p. 16)

Now, assuming that the same sort of identification would have been possible (with some additional effort) for the DNN used by Chang et al. (2018), *ablation* of the

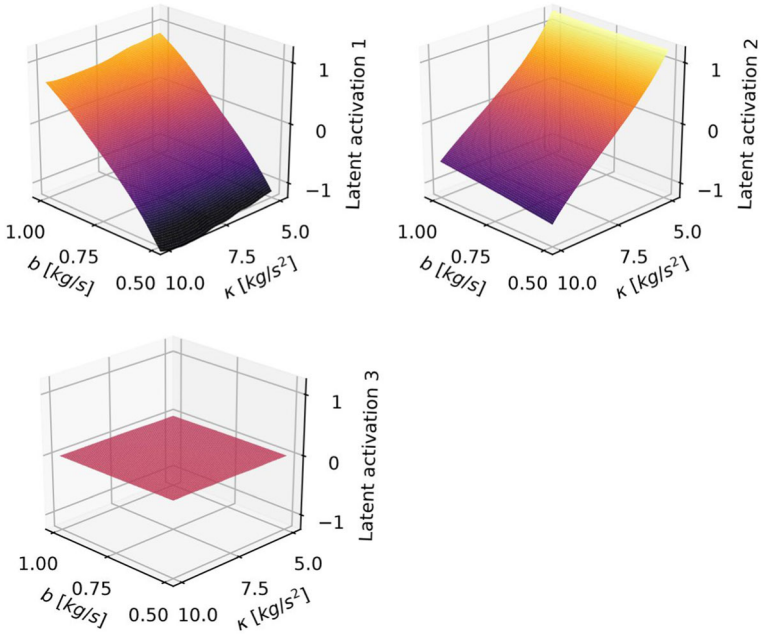

Fig. 4 Activations of SciNet's three latent units when trained to predict the behavior of a pendulum, plotted against the two constants (b and k) describing the damped pendulum equation (both divided by mass). Taken from Iten et al. (2020b), courtesy of the authors. Color available online

respective units in the style of Bau et al. (2018) would have probably led to the exact same losses in performance as the removal of relevant information from the data. In this way, the corresponding activations could have again been shown to function like the respective concepts within the given task. This qualifies them as candidates for functional proxies. (If you wish, you may imagine a particle physicist with Alzheimer's staring blank at a mass histogram for comparison.)

To again supplant the claim of a *mere* proxyhood, though, imagine that the data had been contaminated with artifacts from the data-generation process and that these artifacts correlated with the changes in certain histograms. For instance, particle physicists often use simulated data to train DNNs and it is well known that this can induce spurious correlations with certain *assumed* particle masses (see Kasieczka & Shih, 2020). However, there are of course many further sources of artifact in these complex simulations (Boge & Zeitnitz, 2020, for an overview), and some more subtle such artifacts could perfectly well correlate with the relevant swift changes in histograms *without* thereby correspond to any meaningful representations at all—much like the greyish smudges learned by Bau et al.'s (2018) generative DNN.

In sum, this makes it again entirely reasonable to hold that the DNNs considered here should be said to have learned statistical correlations among numbers, rather than having developed concepts: Because the links to other relevant concepts such as PARTICLE or SCATTER were likely missing, and links to semantically meaningless patterns are to be expected.

Nevertheless, relative to the tasks at hand, the activations learned by the DNNs used by Chang et al. (2018) and Iten et al. (2020a) seem to function the same ways as relevant human concepts would. Hence, it is also entirely reasonable to attribute the *success* of SciNet and Chang et al.'s DNN to *FCPs* for concepts such as MASS, RAPIDITY, DAMPING- and SPRING- CONSTANT.

Let me dispel a final distraction here. It probably goes without saying that finding some sort of parametrized function to describe a data set in an otherwise conceptually rather empty way can *stimulate* the development of new concepts in researchers. For example, the existence of the Rydberg formula $\lambda_R(n, m) = R^{-1}(n^{-2} - m^{-2})^{-1}$, parametrizing the distances between spectral lines, was later claimed by Bohr to have exerted a major influence on his development of the atom model (cf. Duncan & Janssen, 2019, p. 14).

The concepts used by Bohr were, however, fully unknown to Rydberg and Balmer (who devised the predecessor formula), and so the discovery of a parametric function describing an empirical regularity can here be claimed to have stimulated the development of entirely new concepts. The point, however, is not that *researchers* can find new concepts *using* DNNs, but whether DNNs *themselves* can develop concepts, and how that relates to *their* success.

This gives us a first indication as to why DNNs might be *special* in science: It seems plainly nonsensical to claim that a *function* with free parameters, such as the Rydberg function with its adaptable proportionality constant R^{-1}, can develop concepts or even proxies for these. This holds regardless of whether the function is as simple as Rydberg's, or as complicated as a multivariate probability distribution.

What is this intuitive difference in the attributability of self-developed FCPs between DNNs and traditional parametric functions due to? One might think that it resides in the fact that DNNs are implemented on physical machines, but I believe this is a mistake: Having a computer program fit the parameters of a multivariate Gaußian, we would still hesitate to say that it develops even proxies for concepts.

Rather, the difference indeed lies in the fact that DNNs mimic at least some properties of cognitive agents, such as their adaptability across different purposes[19] as well as their 'partial autonomy' (Boge & Grünke, forthcoming, p. 16). More specifically, being 'universal approximators' (e.g. Hornik et al., 1989; Poggio et al., 2020), DNNs can in principle be trained to fit a very wide range of input–output connections, provided they have enough units. This makes it meaningful to even speak of 'learning' here (though this may, of course also be possible in other circumstances), in contrast to the automated fitting of some more restricted parametric function as just considered.

Furthermore, if one uses a loss function that can be meaningfully taken to reflect at least some potential 'aim' of an agent, we may think that a machine equipped with a DNN so trained at least somewhat autonomously navigates its environment.[20] Together with the facts about conceptual interpretability discussed in Sect. 2.4, this makes it plausible that a DNN, in contrast to most other functions scientists use to analyze data, can be associated with the development of FCPs.

[19] Though not yet in the sense of a full domain-general intelligence; cf. Lyre (2020) and the foregoing.

[20] Based on the above remarks on carbon chauvinism and implementation, one may thus of course *speculate* that a machine which realizes some hypercomplex, multimodal DNN will thus eventually satisfy our intuitive criteria for agency, but as I said, I believe indulging in such speculation is presently still premature.

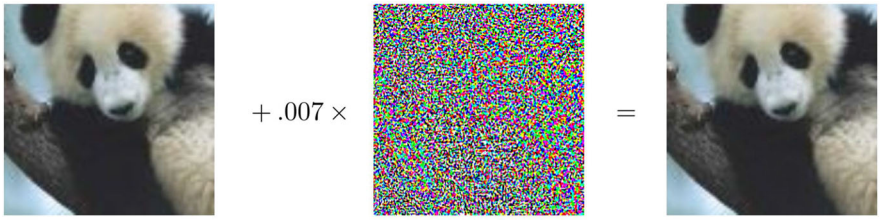

Fig. 5 Famous example of an adversarial, taken from Goodfellow et al. (2014b). Color available online

3.3 The Clever Hans Problem

Despite their remarkable usefulness in science, there is also a well-known problem associated with DNNs and the FCPs they develop, sometimes called the *Clever Hans Problem* (CHP). This is the problem that DNNs often seem to 'reward-hack' (Buckner, 2021) their way through the data by exploiting features that do not generalize beyond training and testing sets, or at least not to all relevant data. The name is instructive: 'Hans' was a horse who appeared to be able to count, and even perform arithmetic operations, using his hoof. But a closer analysis revealed that Hans was looking for subtle clues from his trainer (Johnson, 1911). In other words: In order to respond successfully, Hans reacted to features that did not generalize beyond his training- and testing-situations.

There is a famous set of examples sometimes used to unmask such Clever Hans behaviour (cf. Goodfellow, 2018, pp. 34–35), going by the name of 'adversarials'. In the original sense of the word (Szegedy et al., 2013), these are data-instances (often images) which are perturbed by the addition of noise just enough so as to make the relevant DNN fail (Freiesleben, 2021, for a formal definition). It has been realized, however, that there are also 'natural' adversarials; images of objects shown in a strange pose (Alcorn et al., 2019) or with an unusual background (Hendrycks et al., 2021) that make DNNs fail without having been specifically crafted to do so.

A now-famous example is the image of the panda, originally relatively confidently classified as such by the GoogLeNet DNN, but classified as a gibbon with an even higher confidence by that same DNN after the addition of some random-looking, dedicated noise (Fig. 5). But if a small amount of humanly-invisible noise is sufficient to spoil the DNN's success, it seems that it must have learned to rely on features other than those used by humans to identify pandas.

Note that adversarials could thus possibly be fleshed out as providing further, more direct (though maybe also weaker) evidence of the fact that decidedly semantic properties are likely missing in DNNs. In fact, Fig. 5 already nicely illustrates the relevant intuition, as it seems to suggest that both 'pandas' and 'gibbons' are just decisive pixel patterns to GoogLeNet; patterns that merely *correlate* (imperfectly) with something

humanly *interpretable* in terms of object-level concepts. However, making a somewhat compelling case here clearly requires looking into further technical details again.

There are two recent studies on adversarials I have in mind in particular, introducing DNNs' vulnerability to 'unforeseen attacks' and 'blind spots', respectively (Kang et al., 2019; Narodytska & Kasiviswanathan, 2016; Zhang et al., 2019). Unforeseen attacks are adversarial examples generated in a way that differs from how other adversarials are generated, which the DNN has already learned to master (Kang et al., 2019).

For instance, one may generate adversarials by moving a small step in the direction of greatest change of the loss function along each dimension of an image vector (Goodfellow et al., 2014b). If a classifier vulnerable to these adversarials is then integrated into a GAN architecture in which the generative part produces corresponding adversarials in exactly this way, it can learn to classify them correctly. However, when a *different* kind of method for generating aversarials is used, even an already adversarially trained classifier will likely continue to fail (Kang et al., 2019, p. 4).

Thus, including simulated snow or fog can here lead to severe misclassifications. Furthermore, contrary to what would be expected, heavier snowfall is not generally more likely to produce such misclassifications, even though it more strongly distorts the visibility of the object (ibid.). So the strength of perturbation *for the DNN* does not co-align with the strength of perturbation experienced by a human being. As a matter of fact, even a small number of *randomly added pixels* have been shown to spoil DNNs' successful performance (Narodytska & Kasiviswanathan, 2016), which suggests that whatever the DNN exploits in order to classify correctly is connected to the distribution of the pixels rather than to any 'real' features that could be interpreted as corresponding to concepts.

Blind spots add a second layer to this. These are data instances that are in a sense 'far' from the data encountered during training, while still being correctly classified by a DNN and perfectly well recognizable for human beings (Zhanget al., 2019, p. 5). The sense of distance here is non-trivial but intuitive: Zhanget al. (2019, p. 4) averaged over the k nearest neighbors of training data to the given testing data instance, but in the space spanned by the activations in a hidden layer of some DNN, and with 'nearness' defined by an ℓ_p metric. This basically measures how atypical a given test-image (or other data point) is *for the DNN*. Small shifts and re-scalings of pixels then suffice to create adversarial examples from such blind spots, and prior adversarial training proves ineffective against these (Zhanget al., 2019, p. 6).

Why should any of this suggest that adversarials can be interpreted as providing further, direct evidence of the fact that DNNs learn FCPs rather than outright concepts? Could it not be taken to merely show that DNNs learn concepts that are alien for us? Indeed it could, but as I said, I believe this is quite a stretch.

For, the fact that altering these 'features' in the data sample just a tiny little bit—which implies, exactly, a slight alteration of the classification task, t—can fool the DNN into 'seeing' something completely different should strike us as surprising: It would be like, say, a convicted felon on the run passing a police control by merely putting on a tiny little bit of make up, just because the police officer hadn't seen him in real life before. Explaining such an event by saying that the police officer entertains alien concepts seems fairly contrived. Under these circumstances, we might rather wonder whether said police officer does not suffer from sudden-onset Alzheimer's.

Here is, hence, a different perspective: The activations of hidden units get correlated, during training, with certain distinctive pixel-patterns in images that we, as human beings, recognize as displaying certain objects. The DNN thereby learns to classify them accordingly, based on said patterns and the activations they regularly provoke. However, when we pick images with patterns that have a low correlation with the DNN's activations, we can change these images in a way that hardly makes a difference for us but nullifies *the correlation* entirely. The resulting output is then altered accordingly, for the relevant activations will not contribute anymore. *And that is all that happens*.

To make this just a little more plausible, consider once more the case of Alex. Assume, for the sake of argument, that Alex had only seen perfect circles as round shapes before, but was still able to recognize a yellow ellipse as round. Furthermore, assume that slightly changing the eccentricity of said ellipse and coloring it in a darker shade of yellow made Alex fail. Should we then assume that he had a concept of roundness indeed? Or should we assume that Alex has some sort of *alien* concept? Instead, I believe we should here assume that Alex was only able to group certain compound stimuli into limited equivalence classes—which is not enough for concept possession. Furthermore, I suggest that something of this sort is in evidence when blind spot adversarials (or randomly inserted pixels etc.) make DNNs fail.

The point I am trying to make here could again be challenged on account of a number of observations made by Zhou and Firestone (2019), as well as Buckner (2021), who discuss a range of experiments wherein humans were able to *predict* the erroneous labels the DNN would likely apply to an adversarial. Buckner (2021) argues, accordingly, that our evaluation of DNN performance might be biased. For instance, the limited range of class labels available to the DNN creates a biased evaluation, as the DNN *has* to chose between these labels, and cannot offer a more differentiated account of its reasoning. So maybe the evaluation in the case of blind spots and unforeseen attacks is somehow similarly biased?

I believe that the discussion above demonstrates that the cases considered by Zhou and Firestone (2019) and Buckner (2021) only represent a *selected subset* of all adversarials, which is still somewhat co-aligned with human recognition capabilities. They hence cannot do justice to the full space of possible mismatches between image and classification that can be effected in DNNs by means of adversarial perturbations, and so also do not offer a full explanation of their vulnerabilities (see also Dujmović et al., 2019).

I began this section by claiming that adversarials can be used to unmask Clever Hans behavior, as they show that DNNs do not rely on features they are supposed to, but rather on ones that can easily mislead them. These features have been called 'non-robust' by Ilyas et al. (2019), since small perturbations can destroy them, as we have just seen. Nevertheless, (Ilyas et al., 2019, 1, emph. altered) also hold that "[a]dversarial vulnerability is a direct result of our models' sensitivity to *well-generalizing* features in the data." Thus, while easily perturbed, the features typically exploited by DNNs may in fact be present across broad ranges of available data. This actually leaves open whether they might at least sometimes correspond to distinct hidden patterns that, though not co-aligned with known object- and property-concepts, could be *useful* rather than misleading. I will turn to this question below.

Per se, adversarials fall short of reliably *identifying* the features actually exploited by the DNN though. In this respect, interpretability methods are more promising. Consider, for instance, the approach by Lapuschkin et al. (2019), which builds around a technique called Layer-Wise Relevance Propagation (LRP; Bach et al., 2015), wherein the relevance of a feature encoded into some unit is given by the sum of its normalized contributions to higher-up units, weighted by the relevance of these respective units (for yet higher-up units, and ultimately the output).[21]

There is an obvious resemblance to the works of Bau et al. (2017) or Iten et al. (2020a): Since the first relevance-score is the network output itself, the relevance-scores of previous units are directly connected to their activations and, as in the network dissection method and the study by Iten et al. (2020a), these are then correlated with features of the input. *Unlike* in these studies, however, this is used to generate a relevance map that highlights the features of the *input image* relevant for the overall behavior, not individual units. Hence, this might be seen as providing *indirect* access to DNNs' FCPs, supporting my answer given to reviewer question 5. in Sect. 2.3.

In order to highlight features across data-instance that were relevant to the prediction, Lapuschkin et al. (2019, p. 6) combined LRP with a clustering method. When this method was then applied to a DNN playing pinball, with "excellent results beyond human performance", it could be shown that

> the DNN [...] firstly moves the ball into the vicinity of a high-scoring switch without using the flippers at all, then, secondly, "nudges" the virtual pinball table such that the ball infinitely triggers the switch by passing over it back and forth, without causing a tilt of the pinball table[.] (Lapuschkin et al., 2019, p. 4)

Lapuschkin et al. interpret this behavior as demonstrating that "the model has learned to abuse the "nudging" threshold implemented through the tilting mechanism in the Atari Pinball software." (ibid.) Obviously, this is in a sense a 'valid' solution to the problem of scoring high in a virtual pinball game, and human players might actually be prone to exploit the same sort of mechanism (cf. Buckner, 2021, p. 34). However:

> In a real pinball game, [...] the player would go likely bust since the pinball machinery is programmed to tilt after a few strong movements of the whole physical machine. (Lapuschkin et al., 2019, p. 4)

Phrased in terms of FCPs, the relevant DNN might thus be said to have developed an erroneous FCP of pinball as being mostly a 'nudging game'. This faulty 'conceptualization' would have led to a stark failure when taken out of the comfort-zone of the training and testing data, and into a more realistic implementation of the game.

The CHP is not only a problem for video-gaming and image recognition, though, but also in scientific applications. Particle physicists, for instance, often use simulated data to train ML algorithms, so that they can recognize hypothetical new particles contained in the simulation but not yet recognized in any available real-world data. However, such simulations are built on assumed mass values, and this may, e.g., forestall the identification of particles with *unknown* masses (Kasieczka & Shih, 2020,

[21] Cf. also Samek et al. (2019) for generalizations to this basic rule.

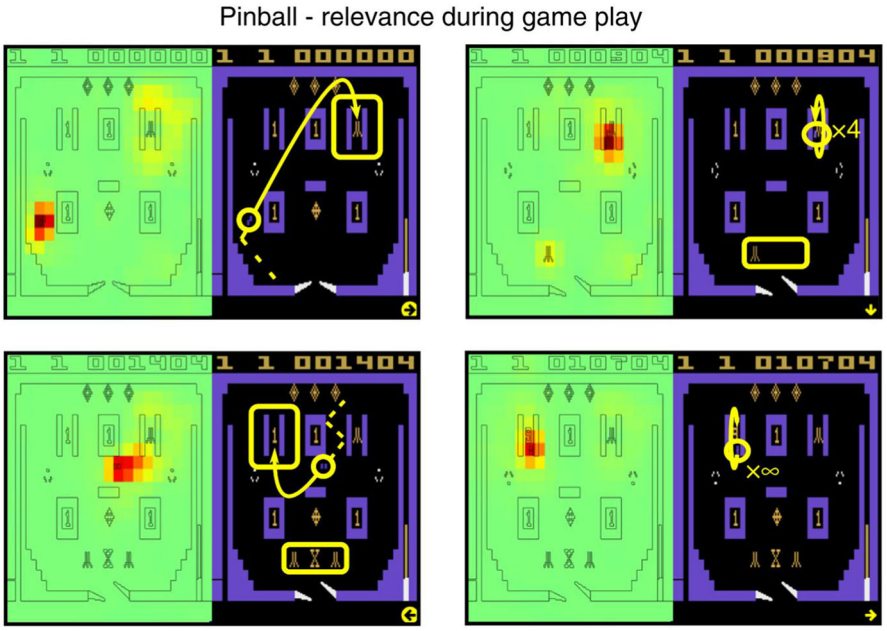

Fig. 6 Gaming behaviour of a DNN viewed through LRP: The DNN focuses on score multipliers, and then uses a nudging mechanism to increase the score indefinitely. The pedals with which to actually flip the ball are not highlighted as relevant at all. Image taken from Lapuschkin et al. (2019) under a CC BY 4.0 license. Color available online

p. 2). Mitigating this issue requires training the DNN on suitable adversarial examples or penalizing it directly by adding a regularizer that penalizes the learned correlation between correct output and mass values (ibid., pp. 2–3).

As can be seen from these examples, the general management of the CHP requires one key ingredient: Insight into the features that the DNN could possibly rely on. In the case of images or video games, this is relatively straightforward, though also not entirely immediate. Figure 6 shows the relevance of certain features in selected frames of the Atari pinball game played by a DNN. It is obvious that the pedals are not relevant at all, but the inference to the DNN's nudging the table in order to succeed requires further information and analysis.

3.4 The Actually Smart Hans Problem

This issue, that the features actually exploited by the DNN usually need to be inferred in more involved procedures than visualization, is less benign than it might seem. Think back to some of the scientific applications discussed above: In the case of particle physics, DNNs could be shown to exploit information on non-linear functions of input variables without being given direct access to the features these represent.

Phrased in terms of FCPs, this means that DNNs are capable of 'conceptualizing' non-obvious features that are highly informative about the underlying physics.

However, there are various exotic concepts around in physics, such as SPIN, COLOR CHARGE, or ENTANGLEMENT, and these were introduced in long-winding back-and-forths between evidence and innovative theorizing. In the words of Susskind (2008, mins. 15:19–15:31): "Nobody has ever understood what the hell Heisenberg was [...] smoking [...] when he invented matrix mechanics."

Now given that DNNs develop FCPs somewhat autonomously and that the discussion surrounding adversarials suggests that they are capable of developing FCPs for features that are "human-inscrutable" (Buckner, 2020, p. 3), it is perfectly conceivable that a DNN develops an FCP that would have to be paralleled by a *novel*, exotic concept, thereby jumping leaps and bounds ahead of the scientific community in its capacity for conceptualizing novel phenomena.

To convince you that this isn't a philosopher's fairy tale, refer once, more to the study by Iten et al. (Sect. 3.2). The simple toy-example of the damped harmonic oscillator clearly adds support to the claim that FCPs exist and are connected to success. However, the main interest of Iten et al. (2020a) was obviously not in proving points about toy examples: Ultimately, SciNet (or some suitable successor) is supposed to convey insight into real data, possibly from highly involved systems, and to elucidate the relevant concepts needed to successfully recover these systems' behaviors. In particular, (Iten et al., 2020a, p. 3) express the hope that "for quantum mechanics," this may aid in "finding conceptually different formulations of the theory with the same predictions". There is continuing interest in this because, despite its current status as the fundamental framework for physics, quantum mechanics faces well-known philosophical difficulties (Boge, 2018, for an overview).

However, given how remarkably difficult it was for physicists to arrive at quantum mechanics and its underlying concepts in the first place, it seems far fetched to hope that the concepts to read off from SciNet's latent units simply jump in one's face. More precisely: In the toy example, extracting these was easy only because it was *already known* against which variables to plot the activations.

The major problem I see associated with DNNs' use in science, then, is that they may develop FCPs based on features that are (a) non-obvious or even "human-inscrutable", (b) present across (vast and complex) data sets, and (c) highly fruitful for scientific prediction and discovery. This will make human researchers fall behind qua being left without the right concepts to (i) comprehend the reasons for the given DNN's success and to (ii) develop theoretical models of their own to advance science in the ways we're used to. This is what I have called the *Actually Smart Hans Problem* (ASHP) above.

What is the detailed connection between FCPs and the ASHP? I believe an analogous problem would arise at a stage where DNNs could more clearly be claimed to have actual concepts. That is, at an imagined future stage where the discussed restrictions that make concept-attribution problematic are absent, we could obviously also face a situation where DNNs learn concepts that equip them with abilities beyond what is humanly possible. The first thing to notice is, hence, that FCPs are *enough* to create the ASHP: DNNs need not even have concepts in order to selectively outsmart us.

Secondly, however, I believe that FCPs also *amplify* the ASHP: Recall that a major reason for rejecting the notion that DNNs have concepts was that they sometimes

connect what seem to be meaningful representations to seemingly meaningless blobs. Hence, the semantic knowledge inherent in a concept is missing in an FCP. Now consider again the imagined future scenario wherein we have much more solid reasons for thinking that DNNs do in fact have concepts. They would then likely also be equipped with communication skills helpful for instructing human researchers on their quest for humanly comprehensible models.

In contrast, so long as DNNs only have FCPs, the 'reasoning' they can offer for certain decisions based on humanly inscrutable features would likely turn up seeming nonsensical to human beings, even though there might be conceptually valuable information hidden in the data, and exploited by the DNN. For example, Google's PaLM has a reasoning prompt wherein it offers a concise reasoning chain for its outputs.[22] However, there is a general worry that the successful examples are cherry picked, and that there will also be nonsensical reasoning chains offered on a larger scale (see Marcus & Davis, 2020).[23] I thus submit that at the present stage, where DNNs arguably only have FCPs instead of concepts, the ASHP is even worse than it would be at a stage where conversations with semantically competent AI could educate us.

To summarize, I agree with Buckner (2018, 2020) and López-Rubio (2020) that it is important to relate DNNs and their present successes to human concepts, as this relation may help us understand said successes better. Furthermore, I agree especially with Buckner (2020) that the non-human feature selection made by DNNs can have major implications for science.

However, as I have explained in detail above, I believe it is premature to associate outright concepts to DNNs. FCPs deliver a more cautious notion that does justice to the controversy over the status of present-day AI. Furthermore, as explained in this section, I believe that this presence of FCPs can make for an actual *problem*, which I have coined the ASHP: Since we arguably desire more from science than mere prediction, successful performance may not be enough. We want to know the 'right reasons' for success; that is, we desire to possess concepts that allow us to parallel our DNNs' successes, thereby giving us *insight* into the information hidden in the data.

Given that the ASHP can arise already now, at a stage where it is at least controversial whether DNNs can indeed be said to have concepts, it may be especially difficult to bypass the fact that we do not have those concepts: If I am right, DNNs themselves do not (yet?) possess systematic, semantic mental representations, and so even reasoning-prompts may prove useless in the kind of situation where the ASHP may arise.

Note that this is a contingent problem; nothing about DNNs strictly necessitates that they succeed in exactly this way. But all the evidence pointing in the direction that DNNs very often 'conceptualize' the data in ways that can be highly fruitful, though at the same time fairly different from our own, clearly suggests that it is entirely *likely* to happen. Furthermore, interpretability methods may partly help, as was demonstrated in the foregoing sections. But when larger numbers of complex, novel concepts are required to reproduce a DNN's success—as could become the case, say, in purported

[22] See https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html; (checked 08/22).

[23] See also https://twitter.com/garymarcus/status/1512067689908375556?s=21&t=P-TTB1D_BllCLGK4-NgZbg0 or https://garymarcus.substack.com/p/what-does-it-mean-when-an-ai-fails (both checked 08/22) in this connection.

replacements of present-day quantum physics—this seems less and less plausible. In other words: The ASHP is a problem that can easily arise already in present-day science, and it might prove fairly hard to overcome.

4 Conclusions

I have argued in this paper that while DNNs may not be literally in possession of concepts, they are able to develop functional proxies for these (FCPs), relative to a given set of tasks, and that this quite likely is the main reason for their success. I have supported both these claims, that to the existence of FCPs and their connection to success, with dedicated empirical evidence from the ML literature, followed by a discussion of a well-known problem associated with the resulting abilities of DNNs to specialize to non-obvious features abstractable from the data: The clever Hans Problem (CHP). This problem is focused on features that are non-robust in the sense that they are easily destroyed by dedicated perturbations and on top of that do not generalize well beyond training and testing data.

However, as I have here argued as well, there is also an opposite, 'Actually Smart Hans Problem' (ASHP): That DNNs could, in virtue of their ability to develop FCPs, specialize to features that *are* well-generalizing while also being highly non-obvious or even human-inscrutable. In virtue of this, DNNs might jump ahead of researchers in their ability to predict complex phenomena in ways that would require novel theoretical understanding of human beings. Actually, several studies from applied ML, as discussed in this paper, seem to suggest that we might be on the verge of this happening.

This problem is only beginning to be recognized in the technical and philosophical literature, and usually in different terms than I have used to phrase it here: Ilyas et al. (2019) define a notion of "useful non-robust features" which are correlated with the desired output in supervised learning but not so when the input is minimally perturbed. And they emphasize that these may be "highly *predictive* features that happen to be non-robust under a *human-selected* notion of similarity" (Ilyas et al., 2019, p. 11; emph. added). But they do not connect this specifically to scientific applications or the possibility that some such DNN-discovered features could require the need for new concepts on the human side.

Similarly, Buckner (2020) suggest several distinctions, including a "cut [that] divides the predictive-but-inscrutable features into artefacts and inherent data patterns detectable only by non-human processing" (ibid., 5), and on top of that (ibid, 3) makes a connection to Goodman's bleen and grue, and so to non-standard concepts. But he displays this more as an opportunity than a problem. Finally, Boge (2021) sketches a challenge that is similar to the ASHP, but is overly specific about the conditions under which this may happen, and rather unspecific about the relation between DNNs and concepts (or FCPs).

I believe that FCPs and their relation to the ASHP give us a fairly clear sense of what is special about DNNs in science: As I have argued in Sect. 3.2, other multivariate methods in statistics (and many simpler ML algorithms) are not associated with FCPs, mostly due to differences in generality, adaptability, in-principle conceptual

interpretability, and partial autonomy. Furthermore, there is a good case that FCPs are responsible, not only for Clever Hans behavior, but also the super-human (or, more generally: unrivaled) performance of DNNs that we are currently witnessing in several domains. Thus, with the current revolution in AI as predominantly brought about by present-day DNNs, we may also be witnessing shift in the way we do science, as they take us a step away from traditional procedures such as the formation of concepts on which we base theories and models to generate successful explanations and predictions.

Acknowledgements I thank audiences at the *Wuppertal-Hannover-Munich Philosophy of Science Network Meeting* and the *Issues in XAI #5: Understanding Black Boxes – Interdisciplinary Perspectives*, as well as Timo Freiesleben, Luis Lopez, and Albert Newen for some helpful additional comments and Michael Krämer and Christian Zeitnitz for some interesting discussions. I also thank four (!) anonymous referees for very helpful suggestions.

Funding Open Access funding enabled and organized by Projekt DEAL. The research for this paper was funded by the German Research Foundation (DFG), largely as part of the research unit *The Epistemology of the Large Hadron Collider* (DFG Grant FOR 2063). Some revisions were also made during FJB's leadership of the Emmy Noether group *UDNN: Scientific Understanding and Deep Neural Networks* (DFG Grant 508844757).

Declarations

Conflict of interest The author(s) declare(s) no conflict of interest.

References

Aad, G., Abajyan, T., Abbott, B., Abdallah, J., Khalek, S. A., Abdelalim, A. A., Aben, R., Abi, B., Abolins, M., AbouZeid, O. S., & Abramowicz, H. (2012). Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B, 716*(1), 1–29.

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

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., & 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* (pp. 4845–4854).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE, 10*(7), e0130140.

Baldi, P. (2021). *Deep learning in science*. Cambridge University Press.

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

Bau, D., Zhou, B., Khosla, A., Oliva, A., & 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., & Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. arXiv Preprint. arXiv:1811.10597

Boden, M. A. (2014). Gofai. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 89–107). Cambridge University Press.

Boge, F. J. (2018). *Quantum mechanics between ontology and epistemology*. Springer.

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*. https://doi.org/10.1007/s11023-021-09569-4

Boge, F. J., & Grünke, P. (forthcoming). Computer simulations, machine learning and the Laplacean demon: Opacity in the case of high energy physics. In M. Resch, A. Kaminski, & P. Gehring (Eds.), *The science and art of simulation II*. Springer. Preprint version from http://philsci-archive.pitt.edu/17637/

Boge, F. J., & Zeitnitz, C. (2020). Polycratic hierarchies and networks: What simulation-modeling at the LHC can teach us about the epistemology of simulation. *Synthese*. https://doi.org/10.1007/s11229-020-02667-3

Branden, C., & Tooze, J. (1999). *Introduction to protein structure* (2nd ed.). Garland Publication.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence, 47*(1–3), 139–159.

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.

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

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature, 588*, 203–204. https://doi.org/10.1038/d41586-020-03348-4

Camp, E. (2009). Putting thoughts to work: Concepts, systematicity, and stimulus-independence. *Philosophy and Phenomenological Research, 78*(2), 275–311.

Carter, N. (2020). *Data science for mathematicians*. CRC Press.

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

Chatrchyan, S., Khachatryan, V., Sirunyan, A. M., Tumasyan, A., Adam, W., Aguilo, E., Bergauer, T., Dragicevic, M., Erö, J., Fabjan, C., & Friedl, M. (2012). Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B, 716*(1), 30–61.

Clark, A. (1993). *Associative engines: Connectionism, concepts, and representational change*. MIT Press.

Davies, M. (2015). Knowledge—Explicit, implicit and tacit: Philosophical aspects. In J. Wright (Ed.), *International encyclopedia of social and behavioral sciences* (2nd ed., pp. 74–90). Elsevier.

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

Dreyfus, H. L. (1992). *What computers still can't do?: A critique of artificial reason*. MIT Press.

Dujmović, M., Malhotra, G., & Bowers, J. (2019). Humans cannot decipher adversarial images: Revisiting Zhou and Firestone. In *2019 Conference on cognitive computational neuroscience*. https://doi.org/10.32470/CCN.2019.1298-0

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

Fazelpour, S., & Thompson, E. (2015). The Kantian brain: Brain dynamics from a neurophenomenological perspective. *Current Opinion in Neurobiology, 31*, 223–229.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge University Press.

Freiesleben, T. (2021). The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines, 32*(1), 77–109.

Freiesleben, T., König, G., Molnar, C., & Tejero-Cantero, A. (2022). Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. arXiv, 2206.05487:[stat.ML]. https://arxiv.org/abs/2206.05487

Goodfellow, I. (2018). Defense against the dark arts: An overview of adversarial example security research and future research directions. arXiv Preprint. arXiv:1806.04169

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014a). Generative adversarial nets. *Advances in Neural Information Processing Systems, 27*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014b). Explaining and harnessing adversarial examples. arXiv Preprint. arXiv:1412.6572

Hand, D. (2019). What is the purpose of statistical modeling? *Harvard Data Science Review, 1*(1), 6. https://doi.org/10.1162/99608f92.4a85af74

Hand, D. J. (2009). Modern statistics: The myth and the magic. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172*(2), 287–306.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15262–15271).

Hinton, G., McClelland, J., & Rumelhart, D. (1986). A general framework for parallel distributed processing. In D. Rumelhart & J. McClelland (Eds.), *Parallel processing* (pp. 45–76). MIT Press.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.

Hutto, D. D., & Myin, E. (2020). Deflating deflationism about mental representation. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), *What are mental representations* (pp. 79–100). Oxford University Press.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & 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., & Renner, R. (2020a). Discovering physical concepts with neural networks. *Physical Review Letters, 124*(1), 010508.

Iten, R., Metger, T., Wilming, H., Del Rio, L., & Renner, R. (2020b). Discovering physical concepts with neural networks: Supplementary materials. *Physical Review Letters*. https://journals.aps.org/prl/supplemental/10.1103/PhysRevLett.124.010508/Supplementary_information.pdf

Johnson, H. M. (1911). *Clever Hans (the horse of Mr. Von Osten): A contribution to experimental, animal, and human psychology*. New York: Henry Holt & Co.

Jones, D., & Thornton, J. (2022). The impact of alphafold2 one year on. *Nature Methods, 19*, 15–20. https://doi.org/10.1038/s41592-021-01365-3

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021a). Highly accurate protein structure prediction with alphafold. *Nature*. https://doi.org/10.1038/s41586-021-03819-2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021b). Supplementary information for: Highly accurate protein structure prediction with alphafold. *Nature Portfolio*. https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-021-03819-2/MediaObjects/41586_2021_3819_MOESM1_ESM.pdf

Kang, D., Sun, Y., Hendrycks, D., Brown, T., & Steinhardt, J. (2019). Testing robustness against unforeseen adversaries. arXiv Preprint. arXiv:1908.08016

Kasieczka, G., & Shih, D. (2020). Robust jet classifiers through distance correlation. *Physical Review Letters, 125*(12), 122001.

Knüsel, B., & Baumberger, C. (2020). Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science, 84*, 46–56.

Krenn, M., Kottmann, J. S., Tischler, N., & Aspuru-Guzik, A. (2021). Conceptual understanding through efficient automated design of quantum optical experiments. *Physical Review X, 11*(3), 031044.

Kriegel, U. (2003). Is intentionality dependent upon consciousness? *Philosophical Studies, 116*(3), 271–307.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking clever Hans predictors and assessing what machines really learn. *Nature Communications, 10*(1), 1–8.

Lehmann, E. (1990). Model specification: The views of Fisher and Neyman, and later developments. *Statistical Science, 5*(2), 160–168.

Levin, J. (2018). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.). Metaphysics Research Lab, Stanford University.

López-Rubio, E. (2020). Throwing light on black boxes: Emergence of visual categories from deep learning. *Synthese*. https://doi.org/10.1007/s11229-020-02700-5

Lyre, H. (2020). The state space of artificial intelligence. *Minds and Machines*. https://doi.org/10.1007/s11023-020-09538-3

Machery, E. (2009). *Doing without concepts*. Oxford University Press.

Marcus, G., & Davis, E. (2020). GPT-3, bloviator: OpenAI's language generator has no idea what it's talking about. MIT Technology Review. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

Mariani, M., Tweneboah, O., & Beccar-Varela, M. (2021). *Data science in theory and practice: Techniques for big data analytics and complex data sets*. Wiley.

McGinn, C. (1988). Consciousness and content. *Proceedings of the British Academy, 76*, 219–23.

Moradi, R., Berangi, R., & Minaei, B. (2020). A survey of regularization strategies for deep models. *Artificial Intelligence Review, 53*(6), 3947–3986.

Narodytska, N. & Kasiviswanathan, S. P. (2016). Simple black-box adversarial perturbations for deep networks. arXiv Preprint. arXiv:1612.06299

Newen, A., & Bartels, A. (2007). Animal minds and the possession of concepts. *Philosophical Psychology, 20*(3), 283–308.

Neyman, J. (1939). On a new class of "contagious" distributions, applicable in entomology and bacteriology. *The Annals of Mathematical Statistics, 10*(1), 35–57.

Orlandi, N. (2020). Representing as coordinating with absence. In J. Smortchkova, K. Doelga, & T. Schlicht (Eds.), *What are mental representations?* (pp. 101–134). Oxford University Press.

Pepperberg, I. (1999). *The Alex studies*. Harvard University Press.

Petti, S., Bhattacharya, N., Rao, R., Dauparas, J., Thomas, N., Zhou, J., Rush, A. M., Koo, P. K., & Ovchinnikov, S. (2021). End-to-end learning of multiple sequence alignments with differentiable Smith-Waterman. *bioRxiv*. https://doi.org/10.1101/2021.10.23.465204

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

Piccinini, G. (2022). Situated neural representations: Solving the problems of content. *Frontiers in Neurorobotics, 16*, 846979.

Piccinini, G., & Scott, S. (2006). Splitting concepts. *Philosophy of Science, 73*(4), 390–409.

Pietsch, W. (2021). *On the epistemology of data science: Conceptual tools for a new inductivism*. Springer International Publishing.

Poggio, T., Banburski, A., & Liao, Q. (2020). Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences, 117*(48), 30039–30045.

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

Rescorla, M. (2020). Reifying representations. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), *What are mental representations?* (pp. 135–177). Oxford University Press.

Ryder, D. (2019). Problems of representation I: Nature and role. In J. Symons & P. Calvo (Eds.), *The Routledge companion to philosophy of psychology* (pp. 233–250). Routledge.

Salmon, D. P. (2012). Loss of semantic knowledge in mild cognitive impairment. *American Journal of Psychiatry, 169*(12), 1226–1229.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer Nature.

Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., & Penedones, H. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706–710.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly, 35*(3), 553–572.

Singer, G. (2021). Conceptualization as a basis for cognition—Human and machine: A missing link to machine understanding and cognitive AI. *Towards Data Science*, September 14. https://towardsdatascience.com/conceptualization-as-a-basis-for-cognition-human-and-machine-345d9e687e3c

Skansi, S. (2018). *Introduction to deep learning: From logical calculus to artificial intelligence*. Springer International Publishing.

Srećković, S., Berber, A., & Filipović, N. (2021). The automated Laplacean demon: How ML challenges our views on prediction and explanation. *Minds and Machines*. https://doi.org/10.1007/s11023-021-09575-6

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

Susskind, L. (2008). *Quantum entanglements, part 1—Lecture 4*. Stanford University. Retrieved April 7, 2021, from https://doi.org/10.5446/15105

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines, 29*(3), 417–40.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

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

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications, 10*(1), 1–9.