Digging deeper with deep learning?

Explanatory understanding and deep neural networks

Claus Beisbart, Institute of Philosophy, University of Bern, Länggassstrasse 49a, CH-3012 Bern, Switzerland, Claus.Beisbart@unibe.ch

Version of June 30th, penultimate draft

To appear in the *European Journal for Philosophy of Science*,

doi:10.1007/s13194-025-00668-y

Abstract. Despite their successes at prediction and classification, deep neural networks (DNNs) are often claimed to fail when it comes to providing any understanding of real-world phenomena. However, recently, some authors have argued that DNNs can provide such understanding. To resolve this controversy, I first examine under which conditions DNNs provide humans with explanatory understanding in a clearly defined sense that refers to a simple setting. I adopt a systematic approach that draws on theories of explanation and explanatory understanding, but avoid dependence on any specific account by developing broad conditions of explanatory understanding that leave space for filling in the details in several alternative ways. I argue that the conditions are difficult to satisfy however these details are filled in. The main problem is that, to provide explanatory understanding in the sense I have defined, a DNN has to contain an explanation, and scientists typically do not know whether it does. Accordingly, they cannot feel committed to the explanation or use it, which means that other conditions of explanatory understanding are not satisfied. Still, in some attenuated senses, the conditions can be fulfilled. To complete my conciliatory project, I further show that my results so far are compatible with using DNNs to infer explanatorily relevant information in a thorough investigation. This is what the more optimistic literature on DNNs has focused on. In sum, then, the significance of DNNs for understanding real-world systems depends on what it means to say that they provide understanding, and on how humans use them.

Keywords: deep neural networks; machine learning; understanding; laws of nature; causal explanation; unification; mechanistic explanation

1. Introduction

During the last few years, researchers have tried to use machine learning (ML), a branch of intelligence (AI), to predict the weather. For example, Weyn et al. (2021) have trained deep neural networks (DNNs) to predict the evolution of four variables (e.g., 2-m temperature) for the two next time steps, on a grid spanning the entire globe. The input comprised data of the same type, collected at two earlier instances, plus some background information on topography and solar radiation. The researchers used several sets of seeds to obtain an ensemble of 32 network models. The resulting predictions were only slightly inferior to those produced by a state-of-the-art model from the European Centre for Medium-Range Weather Forecasts. Astonishingly, almost no scientific knowledge about the weather was used to construct the models. (See Lam et al. (2023) for similar work, and Price et al. (2024) for ML models that even outperformed traditional computer models.)

This is just one example in which ML has proven powerful at prediction or classification. But do ML models also provide understanding? For instance, do the networks trained by Weyn et al. (2021) also help humans to understand why the weather on any particular day does what it does? Many authors are skeptical. Artem Kaznatcheev (2013) writes that “it really is easier to predict than to understand […]. A focus on blind statistics on big data often provides great practical results in the short term (and that is why it wins funding) at the expense of the understanding needed for long term development of science.” Isaac Tamblyn too has expressed skepticism about the idea that ML, and AI more generally, produce understanding: “We’ve certainly developed tools which are faster, and I’ve seen some AI produce results which were not obvious before hand, but I’m not sure I got a general understanding from it” (quoted in Krenn et al., 2022, supplementary information, p. 16).

This skepticism about ML is, in a sense, comforting; it implies that humans remain central when it comes to understanding. The view seems to have a plausible justification, too: ML models are notoriously difficult for humans to understand. They are often described as black boxes, or as opaque (see, e.g., Burrell 2016, Boge 2022, Beer 2023). How, then, can they provide understanding? ML applications are models, and models seem to provide an understanding of a target system only if they are themselves understandable. If this condition is not fulfilled, how can scientists transfer understanding from a model to its target?

In a recent paper, Emily Sullivan (2022) has challenged this line of thought with a more optimistic outlook. She denies that “scientists [are] trading understanding for some other epistemic or pragmatic good when they choose an opaque and complex machine learning model”; instead, she thinks that “model simplicity and transparency are not needed for understanding phenomena” (Sullivan, 2022, p. 110). Although her focus is on whether opacity impedes understanding, part of her argument is supposed to show that some ML models go some way toward providing understanding. Other researchers too have described cases in which, they argue, ML has provided understanding (Jebeile et al., 2020; Knüsel & Baumberger, 2020, Sullivan, 2022, Meskhidze 2023; see also Krenn et al., 2022).

So, can ML models provide understanding or not? The answer depends on what is meant by expressions like “providing understanding,” “enhancing understanding,” or “contributing to understanding.” If the provision of understanding covers any use of DNNs through which scientists infer information about the target system, the correct answer is likely “yes.” Things are different when DNNs are supposed to provide understanding as theories do—by forming an absolutely essential part of an explanation.

This paper aims to clarify the confusing situation in the literature. Focusing on DNNs, I will first focus on one clearly defined sense in which they may provide understanding: I will ask under which conditions humans who learn from DNN output that p is the case have an understanding why p. This question concentrates on a well-defined setting, which I will call the “simple setting” for the purposes of this paper and which is very natural: Scientists can now train DNNs and use them to make reliable predictions about many target systems, without knowing much about the DNNs they are using. I will argue that several additional conditions need to be fulfilled if we are to say that the scientists understand why p is the case based on the fact that the network reproduces p. These conditions are difficult to satisfy, and typically not satisfied in practice. This result helps to explain the misgivings about DNNs, but it is compatible with the idea that DNNs provide understanding in different senses. I will indeed show that DNNs’ use can contribute to understanding in other senses. Since I will not only inquire whether DNNs provide understanding, but also which kind of knowledge or understanding would be required for their providing understanding, my study has consequences for the demands of explainability.

To achieve my purpose, I will undertake a systematic investigation that draws on the recent literature on understanding (e.g., by de Regt & Dieks, 2005, de Regt, 2017, and Hills, 2016), complementing the case-study method employed previously. To provide a robust argument and avoid dependence on the idiosyncrasies of any particular account of understanding, I will adopt an “ecumenical” approach, condensing the previous work on this topic using broad and general conditions that may be interpreted in different ways.

To my knowledge, no work in the recent research literature has done something like this. Boge (2022, Sect. 2.4) comes closest to my project when briefly discussing how DNNs’ opacity impairs understanding of real-world target systems and what is needed to understand a target if one’s understanding is based on DNNs. But his discussion does not consider sufficient conditions of explanatory understanding, is inspired by a specific account of understanding, and is cast in mechanistic terms and thus not ecumenical. Since his crucial example is also focused on interpreting data in terms of their causes, not on fully explaining any phenomenon found in the data, a more extensive treatment of DNNs and explanatory understanding is needed. The present paper is meant to fill this gap. I take its main contribution to be that it provides a systematic, theory-guided discussion of the different ways in which DNNs may, or may not, provide understanding.

I will concentrate on deep neural networks (DNNs), artificial neural networks with a fairly high number of intermediate layers. (See Buckner (2019) for an excellent philosophical introduction to DNNs.) DNNs have proven to be especially powerful but are also opaque. My results extend beyond DNNs and apply to other ML models, but since ML methods form a rather heterogeneous class of models, it is difficult to track the consequences for them in detail. Admittedly, I do have to exclude some DNNs, such as any with an architecture that has been adapted to a specific problem (I will discuss an example in Sect. 4). It is also safer to exclude DNNs that have been given a special construction or training to ensure some explainability (e.g., Liu et al., 2018, Angelov & Soares, 2020). You can think of the DNN under investigation as a garden-variety DNN that scientists have trained for a specific prediction or classification task without investing substantial domain knowledge beyond interpreting the inputs and outputs. To simplify the jargon, I will set classification aside and focus on prediction.

Before I start, a few clarifications are in order. First, my main focus is on understanding real-world phenomena, not on understanding DNNs. Human difficulties in understanding DNNs will play a role in my argument, to the extent that they make it difficult to obtain understanding from DNNs. Still, my primary interest is not in how DNNs’ general interpretability and explainability can be improved (see, e.g., Beisbart & Räz, 2022). Second, my main question is not whether DNNs themselves understand phenomena. I take DNNs to be tools used by humans and ask whether they provide those humans with understanding. (See Tamir and Shech (2023) for the conditions under which practical understanding may be attributed to machines.) Third, my focus is on so-called explanatory understanding (also called understanding why (something is the case), as it is obtained in the sciences.

The paper is organized as follows: In Sect. 2, following a few clarifications about the nature of understanding, I explain my ecumenical approach to explanatory understanding. In Sect. 3, I then systematically address the extent to which DNNs can provide such understanding within the simple setting. In Sect. 4, I engage with objections and consider other senses in which DNNs may contribute to understanding. I draw my conclusions in Sect. 5.

1. Scientific understanding

The question of whether DNNs provide understanding is not entirely clear because the notion of understanding is itself broad and fuzzy. Although understanding is a central goal of science (de Regt, 2020), working scientists often remain silent on what they mean by “understanding.” Even the philosophy of science traditionally neglected the topic (for evidence see, e.g., de Regt, 2017, Sect. 2.1), but this has changed in the last two decades (see Baumberger et al., 2016, for an overview). I can thus clarify the question using valuable insights and distinctions from the recent literature.

First, understanding has often been set aside because it was taken to be subjective (see de Regt, 2017, Sect. 2.1 for evidence). No doubt certain subjective phenomena are associated with understanding, like the aha effect (“Now I got it!”), or the sense of familiarity that people claim to have when they believe they understand something. As Trout (2002, p. 213) puts it, “there is a special kind of intellectual satisfaction—an affective component—that occasions the acceptance of an explanation, a sense that we have achieved understanding of the phenomena.” But Trout warns us that this sense of understanding can be deceptive; it has accompanied “explanations” that explain very little by current standards. Feelings and the sense of understanding will therefore play no part in what follows.

Second, when scientists pride themselves on understanding, they often have in mind their understanding of systems, phenomena, etc. that are independent of human representations. This kind of understanding contrasts with an understanding of language, symbols, and their meaning. Of course, language is itself a part of the world; very often, we can only understand the real world when we understand representations. Nevertheless, for this paper, it is useful to distinguish between understanding representations and understanding some science’s primary target system, which typically will not involve representations (e.g., the climate). My focus here is on understanding real-world target systems.

Third, the literature about understanding follows ordinary language by distinguishing between different kinds of understanding. For our purposes, it suffices to discriminate between *objectual understanding* of some domain of things (e.g., of climate change) and *explanatory understanding* of why or how p is the case, where p is a proposition stating, for example, that a thunderstorm has occurred at some location (for finer-grained distinctions see, e.g., Baumberger at al., 2016). Since I am mainly interested in the sort of understanding science obtains, and explanatory understanding is important there, my focus will be on explanatory understanding. (Sullivan (2022, p. 111) has a similar focus.)

What, then, is explanatory understanding? Sullivan (2022, p. 111) wisely avoids commitment to any particular account of it, of which the recent literature has provided several. Most of these remain neutral on the topic of explanation, so the accounts of explanatory understanding can be multiplied by combining them with various accounts of explanation. However, remaining silent on explanation and its relation to understanding, or merely relying on intuitions about understanding, comes with a price: Arguments about understanding are less likely to carry conviction.

Fortunately, I can escape this dilemma through an “ecumenical” account of explanatory understanding. Drawing on previous philosophical work on understanding, I will formulate general kinds of conditions that can be interpreted in several distinct ways. I will then develop arguments that do not depend on any of these interpretations, and therefore make a robust case for my conclusion. My ecumenical account of understanding is not meant to provide a new explication of explanatory understanding, but a summary of existing ones. It cannot cover every account of understanding that has been proposed, but it should be broad enough to appeal to many authors.

To have explanatory understanding, an agent first needs to have an *explanation*. In my “ecumenical” approach, I take an explanation to be an argument that leads to a statement about the explanandum phenomenon. Many accounts of explanation, such as the DN/IS account (Hempel & Oppenheim, 1945, Hempel, 1965) and the unification account (Friedman, 1974; Kitcher, 1981), think of explanations in this way. Other accounts, such as mechanistic (Machamer et al., 2000) or causal ones (e.g., Lewis, 1986), at least allow that an explanation is an argument.

The reason is that, qua content of a speech act, an explanation consists of language that describes a mechanism or the causes of the explanandum phenomenon. This description has to lead to a statement about this phenomenon, so the explanation can be viewed as an argument, albeit, maybe, a non-deductive or elliptical one. (Salmon’s (1971) statistical account of explanation is one of the few accounts under which an explanation fails to be an argument (p. 77). But it is implausible for precisely this reason.)

Of course, not all arguments are explanations. Considered as arguments, explanations have to satisfy an extra condition. For instance, an explanation will contain a law, describe a salient cause or mechanism, or instantiate a scheme that allows for unification. In an ecumenical spirit, I will leave open what exactly the extra condition is. However, in any event, the argument must not contain irrelevant information. This is a lesson from the irrelevance objection to the DN model (e.g., Kyburg, 1965; see also Salmon, 1990, pp. 49–50).

The conditions for explanatory *understanding*, some philosophers have suggested, parallel those of the traditional JTB-analysis of knowledge (see, e.g., Baumberger et al., 2016; cf. Grimm, 2001; Grimm, 2021, Sect. 2). Accordingly, to have explanatory understanding, an agent must first have an explanation that largely or wholly satisfies some standard of correctness. Maybe it must even be the true explanation; at a minimum, it must be empirically adequate in that it reproduces relevant data about the target phenomenon (cf. the truth condition T). The precise degree to which explanatory understanding has to be based on facts is currently debated under the heading “factivity” (see, e.g., Elgin, 2009; Lawler, 2021); it can be left open for our purposes. Suffice it to say that my correctness demand allows explanations to be highly idealized. I need only exclude explanations that are false or empirically inadequate in all respects, and this exclusion seems plausible.

The agent must also have an appropriate connection to the explanation (cf. the belief condition B). This condition is often supposed to have two parts. (B1) The agent has to accept the explanation in some way: maybe to believe it, or at least to consider it empirically adequate. (B2) They have to be able to work with the explanation—in particular, to make certain inferences on its basis.

Finally, the agent’s acceptance of the explanation must be justified (cf. the justification condition J). An immediate implication of this condition is that my discussion refers to the context of justification. This context is more demanding than that of discovery, in which DNNs’ usefulness is uncontroversial (Duede, 2023).

Conditions paralleling T and B can be found in many accounts of understanding. Think, for instance, of de Regt and Dieks’ account (de Regt & Dieks, 2005; de Regt, 2017; for definiteness, I will refer to the latter). According to this account, a phenomenon P is understood “if and only if there is an explanation of P that is based on an intelligible theory T and conforms to the basic epistemic values of empirical adequacy and internal consistency” (de Regt, 2017, p. 92). (This formulation might look like an explication of objectual understanding, but de Regt is clearly interested in explanatory understanding; see his 2017, p. 2.) A theory, in turn, is intelligible to an agent insofar as they can draw inferences from it without doing exact calculations (p. 102). This account contains conditions that parallel JTB’s T (the theory that gives the explanation has to be empirically adequate) and B (scientists have to be able to draw certain inferences). De Regt does not discuss justification, and does not explicitly require the explanation, or the theory, to be accepted. However, I do not think that he would object to adding these conditions. Admittedly, de Regt argues that models can provide understanding without being believed (p. 134), but my ecumenical approach does not require belief proper; something like acceptance would suffice. De Regt does state that understanding a phenomenon links it to “accepted items of knowledge” (p. 91), and this seems to imply that the explanation is at least partly rooted in what the agent accepts. From a systematic perspective, it would seem odd if an agent understood a phenomenon thanks to a theory or explanation, yet did not accept the theory or explanation in some respect. It is also very natural to require a justification for this attitude.

For another account of explanatory understanding, let us turn to Hills (2016). According to Hills, an agent understands why p is the case if they have both a correct belief that q explains p and a set of abilities. The abilities substantiate the condition that the agent grasps the connection between q and p. Hills requires the following abilities (p. 663):

1. follow some explanation of why p given by someone else.
2. explain why p in your own words.
3. draw the conclusion that p (or that probably p) from the information that q.
4. draw the conclusion that p′ (or that probably p′) from the information that q′ (where p′ and q′ are similar to but not identical to p and q).
5. given the information that p, give the right explanation, q.
6. given the information that p′, give the right explanation, q′.

Some of these abilities, viz. (i), (ii), and (v), encode an understanding of what the explanation says. The explanation is, or can be expressed using, a set of statements; the agent has to be able to understand what these statements say (i) and to formulate explanations in terms of such statements (ii, v). Other abilities go beyond understanding what the explanation says. The agent has to be able to infer p from the explanans q (iii). On top of that, they must be able to draw inferences about counterfactual scenarios—for instance, to infer what would happen if the facts stated in the explanans were slightly different (iv). They must also be able to infer what the correct explanation would be if the explanandum phenomenon differed slightly (vi).

Hills’ account of explanatory understanding includes conditions paralleling T and B. She requires the explanation to be true (T). The agent has to believe that q explains p, and must have related inferential abilities (B). Like de Regt, Hills does not mention justification, and she explicitly does not require an agent to know that q explains p in order to understand why p (her Sec. 4). Still, this does not rule out requiring some justification for the agent’s attitude towards the explanation.

Altogether, then, a brief look at Hills’ and de Regt’s accounts of explanatory understanding has shown that such accounts include conditions paralleling T and B. While conditions paralleling J are rarely mentioned (see Baumberger & Brun, 2016, for an exception), some condition of this kind is plausible. In what follows, I will therefore assume that explanatory understanding can be explicated in terms of conditions paralleling JTB. What we need for explanatory understanding, then, is this:

(E) an explanation—an argument that fulfills an extra condition (e.g., it contains a law of nature) and does not contain irrelevant information;

(T) the explanation’s correctness (truth or empirical adequacy in central respects);

(B1) some acceptance of the explanation (e.g., belief in it);

(B2) some abilities, particularly inferential ones;

(J) justification for accepting the explanation.

Arguably, explanatory understanding comes in degrees, because having the abilities is a matter of degrees (Hills, 2016). Explanations’ informational content can also vary, at least on some accounts (e.g., Lewis, 1973). To simplify my discussion, though, I will gloss over these different degrees of understanding. I will be concerned with a rather low degree of understanding; the exact threshold is implicit in my discussion.

To finish this overview of understanding, it is useful to distinguish explanatory from objectual understanding. According to Kvanvig (2003, p. 192), objectual understanding consists in grasping connections between the parts, or aspects, of the target domain or phenomenon. For instance, to understand ferromagnetism, one has to grasp connections between the relevant order parameter (temperature), the macroscopic signature of ferromagnetism, and the underlying microphysics. This grasp is again a matter of degree (Baumberger, 2019). For Kvanvig, the connections can be of various kinds: for example, probabilistic (certain characteristics are correlated to each other), evidential (one piece of information supports another), logical (certain macroscopic traits follow from the microphysics and bridge principles), or explanatory (the microphysics explains the macroscopic behavior). Now, it is plausible that a grasp of explanatory relationships constitutes explanatory understanding, and therefore that explanatory understanding is part of objectual understanding. Still, some degree of objectual understanding is possible even without any grasp of explanatory relationships; the agent may grasp other relations, such as logical or evidential ones.

With this in mind, I now turn to whether DNNs provide understanding. As I indicated in the introduction, I will initially interpret this question as follows. I assume that a human scientist has used a DNN to obtain a prediction, which I express as proposition p (roughly: such and such an object has property F; or, a certain initial state will be followed by a subsequent state with property F). Here, p need not contain a full description of the subsequent state; it may only describe certain aspects of the latter, e.g., the average temperature of the target system. I assume that the human knows how to train the model and can interpret the output and translate it into proposition p. Training a successful prediction model is for sure not straightforward. In most cases, simply applying an off-the-shelf model to raw data will not produce useful results. It is rather recommended to pre-process the data, and a good understanding of the training data and the general method of training DNNs is pivotal for this step (Karpathy 2019; cf. Vasudevan et al. 2011 for an interesting perspective). Still, the recommended training of DNN involves a couple of steps described in general terms, such as regularization (ibid.). Applying these steps does not require the domain knowledge necessary for explanations. This is plain from the fact that DNNs have been applied to areas in which our explanatory domain knowledge is poor, e.g. in the Deep patient model (Miotto et al. 2016). Note also that the training does not require knowledge of what the DNN does with the data.

For the purposes of this paper, I shall call the setting defined in the last paragraph the “simple setting.” This name is not meant to imply that this setting is very common or a normative standard. Still, for this paper, this setting is interesting for at least two reasons. First, research in the simple setting promises useful results, while not requiring too much prior knowledge. This is evident from general recipes how to train DNNs (e.g. Karpathy 2019). Second, the simple setting can be used to explain the view that DNNs don’t provide explanatory understanding.

My question then is whether, or under which circumstances, the human understands why p is the case. Accordingly, the explanandum is a token event or state, or a token phenomenon. In Sect. 4, I’ll get to the idea that DNNs may explain types of phenomena.[[1]](#footnote-1)

A possible objection is that this question interprets the idea of DNNs’ providing understanding too narrowly; it concentrates on a setting in which scientists do not have explanatory understanding, but for rather trivial reasons. After all, a scientist who makes a prediction by running a computer simulation, not a DNN, may be unable to explain why the predicted phenomenon occurs, simply because they do not know enough about the simulation. This suggests that my investigation of the simple setting is uninformative, because it would lead to negative results about understanding in other cases too.

In response, let me first stress that my simple setting assumes that the scientist knows how to train the network. In the case of computer simulations, an analogous requirement would be that scientists have the basic knowledge needed to build the computer simulation. In most cases, building a computer simulation requires substantial domain knowledge, which often includes laws or causes. DNNs, by contrast, can be trained without domain knowledge. Also, even if the scientist running a simulation does not understand it, the programmers who first developed the simulation did so using domain knowledge, which is often explanatory. Consequently, my interpretation of “providing understanding” would not have problematic consequences for computer simulations, so it does not render this paper’s research question trivial. Also, the simple setting is not meant to exhaust what we want to say about understanding with DNNs. Later in the paper, I will allow for different ways in which DNNs can be said to provide understanding.

1. Understanding phenomena with DNNs in the simple setting
2. The explanation (E)

If DNNs are to provide explanatory understanding in the sense under discussion, they first need to provide an explanation. How might this be achieved? When a DNN infers that p (or a proposition entailing p), this inference is based on the neural network’s input and assumptions made by the network. Thus, an argument leads to p from the input and the assumptions within the network. The natural way to look for an explanation is to ask whether this argument qualifies as an explanation of p. There seems to be no alternative way to associate an explanation with the DNN, so our main question is whether the DNN’s inference explains its conclusion.

As far as prediction is concerned (this point does not apply to classification), here I rely on a well-known connection between explanation and prediction: An explanation includes a belated prediction (cf. Hempel & Oppenheim, 1948, p. 138). However, it would be too simplistic to equate every prediction with an explanation. As Scriven (1962, p. 177) has rightly argued, a prediction is a description of a future event; an explanation, qua argument, is more than that. Further, to be an explanation, the argument has to satisfy an extra condition (e.g., to contain a law of nature among its premises). Still, it is reasonable to claim that an explanation would allow scientists to predict the explanandum phenomenon (or its likelihood) if the latter was not yet known. DNNs that cast predictions satisfy this condition.

But are the other conditions on explanation fulfilled? In particular, is one of the extra conditions fulfilled—for example, does the argument contain a law?

The problem is that we cannot tell whether any particular one of the extra conditions is fulfilled. Without knowing more about the DNN, we cannot tell whether its inference is based upon laws of nature, whether it specifies causes or a mechanism, or whether the inference instantiates a scheme that provides unification. Admittedly, we cannot exclude that one or the other condition is fulfilled, but we don’t know it to be fulfilled.

To show this in more detail, it is useful to reconstruct a DNN’s inference as an argument. The input provides content that yields a premise. Strictly speaking, the input consists of numbers, but we can translate it into propositions stating that certain characteristics (temperature, air pressure, color, …) have these values. Call the related proposition i.

The DNN itself can be regarded as a function *f* that maps input to output. Thus, the DNN contains (at least) the assumption that the claims encoded in the output depend on the claims encoded in the input as *f* has it. Call this proposition d (for DNN).

I therefore arrive at the following argument:

(i) What the input contains about the target system is true.

(d) The target system behaves such that the function *f*,realized by the DNN, correctly maps input information to output information about the target system.

(r) Thus, what the output contains about the target system is true.

For any concrete input, the argument can be recast as follows:

(i′) At an initial time *t0*, the characteristics X, considered in the input, take the values x in the target system.

(d′) For a later time *t*, the characteristics Y, specified in the output, depend on the values of X at *t0* as specified by the function *f*,realized by the DNN.

(r′) Thus, at *t*, the characteristics Y take the values y in the target system.

Here x and y are vectors of concrete values, respectively, of the input and output variables X and Y. (r′) is or entails p.

If the prediction is meant to be conditional, the general argument can also be reconstructed as follows:

(d)

Thus, if (i) then (r).

Neither of these arguments manifestly fulfills any of the conditions that distinguish explanations from non-explanatory arguments. (d) does not coincide with a known law of nature (or an immediate consequence of one), it does not give causes nor does it provide a straight-forward description of a mechanism. Admittedly, (d) may be decomposable into a set of assumptions, one of which is a law of nature, refers to variables that are causally relevant to the explanandum phenomenon or describes a mechanism. But unless more about the network is known, we cannot tell whether this is true. As far as the unification account is concerned, the argument from (i) to (r) does not manifestly instantiate an argument scheme that leads to unification. For this to be the case, that argument scheme would have to be instantiated in many explanations, or be similar to argument schemes that are used in many explanations. While a DNN can be used for a range of inputs, its scope is typically quite narrow. For instance, the DNNs trained by Weyn et al. (2021) predict the weather for a broad range of initial conditions—but there is no reason to assume that they can also predict other phenomena (e.g., from fluid dynamics), and it is unknown whether similar networks can predict the behavior of other systems. By contrast, a good explanation of the weather depends on the Navier-Stokes equations, which can also be used for completely different systems. Accordingly, explanations in terms of the Navier-Stokes equations are associated with argument schemes that work for completely different kinds of systems. Of course, the workings of any DNN can be cast as an argument of the form given above, where only the function *f* in premise (d) differs, but this similarity is too superficial to provide the unification achieved elsewhere in science.

At this point, it is illuminating to compare DNNs to computer simulations. The latter can also be regarded as arguments (Beisbart, 2012), but in comparison with DNNs, their underlying arguments can be cast in a more informative way. This is because scientists have invested substantial assumptions in the simulations. For instance, they assume that the air in the atmosphere obeys the Navier-Stokes equations (or a specific approximation). Accordingly, we can say that these computer simulations use the Navier-Stokes equations as premises. The Navier-Stokes equations are laws of nature, or at least immediate consequences of such laws (viz. conservation laws). They are also used in many other explanations. Accordingly, qua arguments, computer simulations contain laws of nature, or assumptions, that also figure in other explanations. Hence, an extra condition of explanation is fulfilled. No analogous point holds for DNNs, unless we succeed in describing them in a more illuminating way. The challenge is to decompose premise (d) into assumptions that describe what the DNN does such that these assumptions (or, maybe, a proper subset of them) satisfy the extra conditions on explanation. The challenge may well be met—indeed, so far, I have only proposed a minimalist reconstruction of a DNN in terms of an argument. Still, to move beyond such minimalist reconstruction, one must know more about a DNN than how to train and run it successfully. One must interpret the DNN using characteristics of the target system. This is difficult because DNNs are most often described using their nodes, the neurons, and their mutual connections. Depending on the details, we may say that the nodes of a layer achieve a convolution or max-pooling, and accordingly that the network engages in abstraction (Buckner 2019), but this is very general and doesn’t entail specific assumptions about the target system that may be explanatory.

Before I continue, a note on terminology. I say that a DNN contains an explanation if, and only if, its work (i.e., the second premise above) is properly spelled out using assumptions that fulfill an extra condition on explanation. Either at least one assumption in the reconstruction is a law, provides causal information, describes a mechanism, or the assumptions belong to an argument scheme that provides unification. If any of these conditions is fulfilled then the set of assumptions contains a crucial part of the explanans (the other part of the explanans is implicit in the input, which the first premise represents). However, for convenience, I shall say that the DNN contains the explanation. (I allow that the set of assumptions characterizing the DNN may contain a few more assumptions than are needed for the explanation, such as assumptions that project a 3D state description as a 2D image.)

Admittedly, this definition of “a DNN containing an explanation” is a little vague because the idea that the working of a DNN is properly spelled out using assumptions is itself not entirely clear. A set of assumptions about the target phenomenon unequivocally spell out the working of the DNN if they describe functional dependencies between some nodes’ activations. But we must also allow for the possibility that a set of assumptions properly describe the activity of a DNN because they represent that activity at a higher level. The general problem is that it is difficult to describe what the DNN does by referring to properties within the target system. Also, even if a DNN does not contain a full explanation in the sense described, it may contain explanatory information. For example, it may contain some assumptions about a mechanism without fully characterizing it.

But don’t researchers sometimes know that DNNs contain explanations in the sense I just described? In a recent study, Lemos et al. (2023, p. 1) have presented “an approach for using machine learning to automatically discover the governing equations and unknown properties (in this case, masses) of real physical systems from observations.” To do this, the researchers trained a network using observational data about the orbits of the planets in the solar system. They then used symbolic regression to show that the network implemented Newton’s law of gravitation. Surely, then, in this example, the network contained the explanation of those phenomena that it was able to predict (the orbits).

On closer analysis, however, this example of a DNN is rather peculiar because the scientists invested extensive knowledge about the explanation to set it up: They used a special network based upon a graph, the edges of which were known to represent forces between two bodies, and used Newton’s second and third laws as assumptions. Consequently, the network only recovered the values of the gravitational forces once given this framework of Newtonian mechanics. This is still a significant achievement, since it involves a multi-body problem with many mutual interactions. Nevertheless, the example can aptly be described as “explanation in, explanation out”: Scientists can only claim that the DNN contains the explanation because its construction was based on explanatory knowledge in the first place.

Even if a DNN contains the relevant laws in the sense defined above, only human interpretations make this the case. Lemos et al. (2023) applied a symbolic regression to check whether the DNN had learned Newton’s law of gravitation. What it showed, effectively, was only that certain variables in the network took (to an excellent approximation) the values of the Newtonian gravitational forces, in suitable units. However, this does not show that the network recognized these values as the values *of forces*. Similarly, there are algorithms that can identify causal graphs and determine the strengths of treatment effects (Buijsman, 2023), but it is only human attribution that identifies certain variables as causally relevant to others.[[2]](#footnote-2)

All in all, despite the results Lemos et al. (2023) obtained, it remains true that most DNNs provide no manifest explanation of their results. They may contain explanations, but some efforts are needed to find out whether this is the case.

Things become even worse when we recall that many explanations are expected to explain phenomena, not just data (see Bogen & Woodward, 1988, for the distinction). For instance, scientists may want to explain why a thunderstorm has occurred, and are using a DNN like the one Weyn et al. (2021) trained. What it infers is, strictly speaking, not a statement to the effect that a thunderstorm has occurred. Instead, the DNN outputs lists of numbers resembling data in which the thunderstorm manifests itself as a pattern. Not everything in the output is part of this pattern and necessary for a thunderstorm; the related pattern can be instantiated in many ways. Accordingly, the argument running from the DNN’s input and assumptions to the result contains a lot of information that is irrelevant to the explanandum phenomenon. This suggests that an important condition on explanation is violated because there is too much irrelevant information; the DNN’s inference does not work at the level at which the phenomena are described. Note, though, that this point does not apply to all DNNs and all putative explanations. First, some DNNs can infer that a thunderstorm has occurred (for instance, they may be, or contain, classifiers for which “thunderstorm” is a label). Second, sometimes scientists want to explain why a certain state has occurred, and the state description encoded in a DNN’s output may not be redundant.[[3]](#footnote-3)

The result, then, is that, at best, we cannot tell whether the DNN contains an explanation. Despite this negative result, let us look at the other conditions of explanatory understanding and see to what extent they may be fulfilled, maybe in an attenuated sense.

1. Attitude (B1), justification (J), and correctness (T)

The conditions that refer to the agent’s (here, a scientist’s) attitude (B1) and an explanation’s justification (J) can be treated together. This is because, typically, agents will have a certain attitude toward an explanation only if they have a justification for accepting it.

In the previous section, I argued that one cannot tell whether the DNN contains an explanation unless one knows quite a lot about the DNN. In my argument, I have often appealed to scientists’ lack of knowledge, but the question ultimately concerned whether the DNN contains an explanation, not whether scientists know this. In this section, I can re-use what I have said about scientists to make assertions about their mental states. If they have merely trained the network and successfully used it, they do not know whether it contains an explanation, nor do they know the explanation itself. (They may have an independent explanation of the result, but this explanation would not be provided by the network.) Accordingly, even if the DNN does contain an explanation, it is not an explanation that they can justifiably believe.

Still, let us try harder to make sense of the idea that DNNs can provide understanding. Maybe B and J can be fulfilled in an attenuated sense. There are propositions about the DNN that scientists can *accept.* They can, first, accept the premises (i) and (d) from the above argument, and that these lead to the conclusion. They can, second, also accept that this argument is the correct explanation (which is equivalent to saying that the network contains the correct explanation; call this claim ARG-EXP). As I argued above, the argument does not manifestly qualify as an explanation, but it is possible to think that the argument can be decomposed into an explanation. In some attenuated sense, then, scientists can think that the explanation is correct and therefore accept it. Things are similar in a case where I trust that Berta knows where the party is, although I don’t know where it is.

What scientists come to accept about the DNN can even have some *justification* for them. Consider first the argument associated with a DNN’s prediction, moving from (i) and (d) to (r). There can be good reasons to believe that not just premise (i), but also its crucial premise, the assumption associated with the DNN (d), is empirically adequate or true. If the DNN has a good track record of successful predictions regarding observable features of the target system, this is evidence of empirical adequacy and even that the function realized by the network correctly maps input characteristics to output characteristics. (Input and output characteristics need not only be observables.)

There is a problem, however, if some of a DNN’s results have proven incorrect, as often happens. One may then think that the minimal assumption contained in the DNN has been falsified. Accordingly, there would be reasons to take the explanation within the DNN to be incorrect. In what follows, I will abstract from this problem and assume that the DNN has not failed. One reason for this assumption is that I want to investigate how far we can get if we want to claim that DNNs provide understanding. For another thing, some of the failures may be disputed if, arguably, the input does not belong to the DNN’s intended applications. Finally, an argument forming an inductive-statistical (IS) explanation does not always produce the correct result.

Consider second the claim that the DNN contains the correct explanation (ARG-EXP). If the network is known to work with variables that scientists take to figure in the relevant laws, to be causally relevant, to describe a mechanism underlying the explanandum phenomenon, or to figure in argument schemes that achieve unification, then there are reasons to think that the DNN contains an explanation. These variables may be part of either the input or the output. For instance, in the example of weather prediction, the input variables temperature and air pressure at some earlier time are explanatorily relevant to the temperature and air pressure at a later time. We are concerned with a causal story that unfolds on a certain level of description, traced by input and output. In other examples, some part of the input may characterize one level while the target phenomenon is described at another, so this part of the input may contain variables that describe a relevant underlying mechanism.

There may also be reasons to think that the DNN works with variables that figure in laws, causal stories, or mechanisms, even if these variables are not part of the input or output. The study by Lemos et al. (2023) is a case in point. Here, and in other cases, scientists can show that certain variables, calculated by nodes in the inner layers of the network, figure in laws, trace causes, etc. Note, though, that it does not suffice if the network contains variables that are taken to be explanatory for one reason or another; the DNN must also use them to make its predictions. Otherwise, it is difficult to reconstruct the working of the DNN with an argument in which this variable figures prominently. Still, research on explainable AI employs certain methods that can detect which variables or nodes have an impact on the result, or on the quality of predictions (e.g., Meyes et al., 2019).

If there are reasons to think that a DNN contains and uses explanatory information, there can also be reasons to think that its explanation is correct. If the DNN makes correct (conditional) predictions for some range of inputs, scientists can reason that the explanatory information within the DNN must be correct. It would be a surprise if a DNN could successfully predict or classify anything by using false explanatory information.

This reasoning is an inference to the best explanation (of the DNN’s success, not of the target phenomenon). This type of inference is generally controversial, but this is not the place to discuss that. There is a more specific problem, however, because there may be an alternative explanation of a DNN’s success—for example, that the network contains a mere how-possibly explanation, not the correct how-actually explanation. (See Grüne-Yanoff (2009) for similar problems in the context of computer simulations.) Evidence would be needed that the explanation within the DNN is more than a how-possibly explanation. Since this is a well-known challenge that is not specific to DNNs, I will not elaborate on it.

In sum, the B1 and the J conditions are difficult to fulfill because DNNs do not manifestly contain explanations that scientists may accept. Still, scientists can accept the minimal assumption a DNN contains, and they can accept that the DNN contains the correct explanation. There can be reasons to accept this, in particular, if the scientists know that the DNN uses the values of variables that figure in relevant explanations. But, as before, this is a bit “explanation in, explanation out”: The DNN is known to be explanatory only if prior knowledge about the explanation is available.

Still, is it not possible to infer that a DNN contains an explanation, without drawing on prior explanatory knowledge? Suppose that a DNN makes successful predictions in an extremely broad range of cases. It then seems plausible that the DNN must trace relevant laws or causes, or an underlying mechanism. What kind of explanation the DNN gives remains open—it might be a complicated explanation at the level at which the target phenomenon is described, or an explanation in terms of an underlying mechanism. However, the thought is that the DNN could not be so successful if it were not sensitive to the correct explanation. A similar conclusion suggests itself for the unification account of explanation: If the DNN’s range of applications is extremely broad, then it does achieve unification.

I do not have a principled objection against this line of reasoning, but I think that it is inappropriate for present-day DNNs, which typically have a small range of applications. Accordingly, they may be able to make successful predictions simply by picking up on correlations, without using laws, etc.; and the unifying power is too small. It is difficult to say whether this kind of reasoning might at some point become legitimate, and I will leave this question for another opportunity.

So far, my discussion in this section has focused on the conditions B1 and J, and on whether they can be fulfilled in an attenuated sense. By answering this question, I have implicitly also addressed whether condition T may be fulfilled in an attenuated sense. This is because we can address T only via reasons that speak in favor of it, which I have discussed in relation to J. Accordingly, there is no need to say more about T; there can be reasons to think that the network contains the correct explanation.

1. Abilities on the agent’s part (B2)

I now turn to the remaining condition: the abilities. To have understanding, an agent must be able to make certain inferences. For de Regt (2017), these inferences draw consequences from a theory in a qualitative way. For Hills (2016), they concern the explanans, the explanandum phenomenon, and modifications of both.

There is an obvious problem if a scientist is supposed to make these inferences: The scientist does not have an explanation at hand that may be used as their basis. Unless the network is very well understood, at best the scientist can reason that there is an explanation, but cannot state that explanation, and therefore cannot use it to draw inferences. Furthermore, even if a scientist knows what explanation a DNN contains, they probably cannot use it to draw inferences; since DNNs contain an enormous number of parameters, the explanation is likely to be too complicated for that. In this respect, DNNs differ from good theories, which are more parsimonious with their parameters.

Still, let us look harder for a way in which scientists may draw the required inferences, if only in an attenuated sense. I can think of two possibilities.

One possibility is that a scientist gains an understanding of what the network does, and can then anticipate, in a qualitative way, how the network will react to certain inputs, or how the output changes if the input is changed. For instance, perhaps the scientist has obtained a sense for what kinds of input lead the network to predict a thunderstorm. Given the assumption that the DNN is empirically adequate, the scientist can use their understanding of the network to make inferences about the target system, as Hills requires.

There are several ways in which an agent may obtain this kind of understanding of a network. For instance, a scientist may have a model of what a certain DNN does. Kuorikoski (2011, Sec. 4.1) explores this option when discussing computer simulations. Alternatively, by running the DNN many times, the scientist may gain the ability to anticipate its qualitative results intuitively, without recourse to an explicit model. Lenhard (2019, p. 100) considers an analogous acquisition of an inferential ability for computer simulations. We may then say that the scientist has been trained to obtain the DNN’s results. Their anticipations of the DNN’s outputs may qualify as expert judgments (cf. Majszak & Jebeile, 2023, for expert judgment in climate science; see Schubbach, 2019, for a comparison with AI).

Let us grant, then, that the scientist has obtained sufficient knowledge of the DNN’s behavior to make some of the inferences that an understanding of the real-world target system requires.[[4]](#footnote-4) Would these abilities be sufficient to fulfill condition B2? This is doubtful. The problem is that the inferences are not built upon an explanation known to the scientist. Even if the DNN does contain an explanation of the processes in the target system, the scientist does not have this explanation at hand. Accordingly, it is more appropriate to say that a scientist who can make the required inferences has some degree of objectual understanding of the phenomenon that is predicted by the DNN. However, there is no reason to count it as explanatory understanding because it is not based upon an explanation. In this regard, I should note that Hills’ account requires the ability to understand the explanation itself. A scientist cannot have this ability without knowing an explanation that is implicit in a DNN.

There is another possibility to grant the scientist the required inferential abilities, viz. to allow them to use the DNN itself (Kuorikoski (2011, Sect. 4.2) makes the same point about computer simulations). By using the DNN, the scientist can “infer” what the DNN does. This seems to parallel de Regt’s account, in which a scientist uses a theory in order to discover its qualitative consequences. By using the DNN, the scientist can also make inferences about counterfactuals in which the explanans and the explanandum phenomenon are modified, as Hills requires. For instance, the scientist can run the DNN on a modified input to find out how the output changes.

Discussing this possibility is reminiscent of a point that Teller (1980) has made against Tymockzo’s (1979) claim that humans cannot survey the inferences a computer has made in a computer-aided proof. Teller argues that a human can do this with the computer’s help. Similarly, why not allow that the scientist uses the DNN to make the inferences required for understanding?

Once more, though, it is doubtful whether the ability to use the DNN to make inferences is sufficient to fulfill condition B2. First, the question of whether DNNs provide understanding is naturally taken to ask whether DNNs provide *humans* with understanding, where these humans do not rely on further help. A significant part of the literature on understanding DNNs seems to take this perspective. For instance, the claim that DNNs are opaque is typically understood to say that unaided humans cannot easily know or understand the workings of DNNs (e.g., Tymoczko, 1979). This perspective is very natural, since we care about whatever understanding we have on our own. If computers (DNNs) are used to make the required inferences, what has understanding, strictly speaking, is only the coupled system consisting of ourselves and the computer (see Clark & Chalmers, 1998, for this idea). This coupled system is another subject, not the one in which we were initially interested. A human who is trying to obtain explanatory understanding seems not to have it if they must use a computer to make inferences about the target system. Second, there is the problem we have noted before: The DNN may not contain any explanation at all. (If it does, however, we may say that the coupled system draws on it.)

In sum, scientists who use a DNN to classify or predict phenomena typically cannot make the inferences that explanatory understanding would require. The main reason is that they have no explanation at hand that they could use to make these inferences. I have explored possibilities in which scientists obtain the ability to make some of the required inferences. However, these possibilities are not promising because in each case the inferences are decoupled from the explanation.

Altogether, this section’s results are sobering. The suggestion that DNNs provide understanding in the simple setting faces the fundamental problem that we (or the working scientists) do not know whether the network contains an explanation at all. True, DNNs make inferences, and their working can be described using an argument. However, common descriptions of DNNs (e.g., as data models or as models of neuronal networks) do not entail that the argument fulfills the extra conditions which standard accounts of explanation require. Consequently, scientists working with DNNs typically do not have explanations at their disposal, and this makes it rather difficult to satisfy the JTB-like conditions for understanding. Unless more about the network is known, there simply is no formulated explanation that scientists can accept, and the issue of justification does not arise. Scientists can only accept that the explanation exists within the network; while they may, in some sense, be able to make some of the inferences which understanding requires, they cannot use the explanation to do so.

Still, this result does not imply that DNNs cannot provide understanding in the simple setting. If rather strong conditions are fulfilled, a scientist who has reproduced p using a DNN can be said to understand why p. The central conditions hold that the DNN contains an explanation of why p, and that the scientists legitimately accept this explanation and can work with it. This requires them to have some understanding of the DNN: They have to know that the network is sensitive to explanatorily relevant characteristics. In this sense, the network has to be interpretable. Further, the scientists must commit themselves to accepting the explanation and have reasons for this. The success of the DNN in prediction tasks may give the scientists some such reasons.

If a DNN does not provide a scientist with explanatory understanding, it can still go some way toward this. For instance, a scientist may have good reasons to think that the DNN contains an explanation. Maybe, they have used an interpretability method to infer some aspects of the explanation. They may further think that the explanation in the DNN is correct and have reasons for this. Finally, they may also use the DNN to evaluate counterfactuals and thus do something that can be done with explanations. This doesn’t imply full explanatory understanding but at least a mental state that is somehow reminiscent of such understanding.

My results are in rough agreement with Boge’s (2022) claims about DNNs and understanding. Boge first argues that the (w-)opacity of DNNs impairs understanding because their model variables and parameters lack content about the target system. I agree with Boge that these variables and parameters are most easily interpreted in terms of neurons, but it remains possible that they contain explanatory information about the target. Boge then suggests that DNNs might improve understanding if (I) their inputs and outputs are understood, (II) it is clear what a DNN represents, and (III) the represented content has been mapped to underlying mechanisms (in his example, the mechanism underlying the data, but I have suggested that this is not the right target). Boge’s conditions seem to be on the right track, but on closer inspection, it is not clear what (II) means (in his case study, knowledge of what a DNN represents is based only on training and does not refer to an explanation). (III) is cast in mechanistic terms and is not sufficient, even if generalized to other types of explanations: As I have argued, the DNN must also use the explanation identified within it to obtain its prediction. Finally, Boge neglects the scientist’s attitude and abilities, and is not explicit about justification.

My results are rather sobering, then, but are they true? In what follows, I discuss possible objections. Some of them are motivated by other philosophers’ more optimistic results, allowing me to engage with the recent literature and to investigate how DNNs might contribute to understanding in a more positive way.

1. Discussion

According to a first objection, DNNs do contain or provide explanations, but of a new kind—explanation sui generis. This claim may be supported by the observation that an all-encompassing philosophical theory of explanation has proven elusive. Many philosophers, such as de Regt (2017) or Mantzavinos (2016), embrace a pluralist view that allows for several irreducible kinds of explanations. Why not say, then, that DNNs introduce a new kind? By contrast, in this paper I have tried to relate DNNs to well-known types of explanations. This may have been mistaken, since the desiderata that make an explanation good are subject to change. The idea that new technologies enable new kinds of explanation is nicely captured by Epstein and Axtell (1996, p. 20):

Perhaps one day people will interpret the question, “Can you explain it?” as asking “Can you grow [simulate] it?”

In a similar way, scientists may come to accept DNNs as a source of explanations. However, I do not think that they will, or should. Introducing sui generis explanations would violate the principle of parsimony. Further, it is commonly accepted that explanation is a special kind of achievement that differs from others, like mere description. This distinctness is difficult to maintain if anything that reproduces an explanandum phenomenon qualifies as an explanation. Even if explanations do not share a common essence, they must be connected via something like family resemblances. Accordingly, explanations must involve something like laws, causes, or mechanisms, or they must provide unification. An ingredient or feature of this kind must demonstrably be implicit in a DNN if that DNN is to provide understanding. And to show this requires more work than running the DNN successfully in a range of applications.

The second objection is a variation of the first. It claims that DNNs do explain, not in a sui generis way, but qua model. After all, it is often claimed that models can explain phenomena (e.g., Frigg & Hartmann, 2020, Sect. 3.3). DNNs can be considered to be models; seen in this way, they may be able to explain, whether or not they are known to contain explanations as I defined them earlier.

However, it is doubtful that DNNs are models of a kind that can explain the phenomena within a target system. (Of course, DNNs may contain explanations, but the idea of the objection is that DNNs explain qua models, whether or not they are known to contain explanations.) Leaving aside to what extent model-based explanation is an independent form of explanation (Lawler & Sullivan, 2021), I will focus on problems specific to DNNs. As Boge (2022, pp. 45–46, 52) has clarified, DNNs can be conceptualized as models in several distinct ways. First, they can be regarded as models of a network, or as very simple models of brain tissue consisting of interconnected neurons. Second, they can be considered as models of human learning. However, either way, they are not models of their target systems (unless the target system happens to be a network or a human).

Third, DNNs can be thought of as models that represent patterns within data in an illuminating way – as models close to data models (see, e.g., Frigg & Hartmann, 2020 for this notion). However, such models do not provide explanatory understanding. One cannot explain the distribution of human heights by discovering certain patterns in the distribution or by fitting a Gaussian to it. Independently of Boge’s threefold distinction, DNNs as such are not the right kind of models to provide explanatory understanding. When models explain phenomena and provide understanding of them, this is either because they contain laws, mechanisms, and causes, or because they are accessible, perhaps even intuitive, and connected with stories (cf. Hartmann, 1999). But scientists typically do not know that a DNN fulfills one of these conditions.

A third objection draws on recent work on understanding phenomena with ML or DNNs. Some authors have argued that machine learning can lead to understanding, so this strand of the literature might seem to contradict my results. As I will show, it contains valid examples in which DNNs provide understanding in senses unlike the one I have considered so far. By articulating these senses, I can reconcile skepticism about DNNs’ power to provide understanding with case studies that show how DNNs can contribute to understanding.

Jebeile et al. (2020) consider DNNs that are used within larger climate models (Gentine et al., 2018). Roughly speaking, in their example, DNNs are used to model certain aspects of convection. They are trained with outputs of modules of computer simulations and then used in place of these modules. Jebeile et al. (2020) consider five desiderata that they take to foster understanding: intelligibility, representational accuracy, empirical accuracy, consistency with known physics, and the possibility to specify the domain of validity. They claim that DNNs (and other models) lie on “the same continuum where these various criteria of understanding come in degrees, and that therefore machine learning methods do not necessarily constitute a radical departure from standard statistical tools, as far as understanding is concerned” (p. 1877).

The results that Jebeile et al. obtain about their case studies are plausible. Although their results sound more positive about DNNs’ contribution to understanding, they are fully compatible with my claims so far. This is because, strictly speaking, Jebeile et al. study to what extent a climate model *incorporating* a DNN can provide understanding; they are interested in DNNs’ contribution to understanding as part of a larger model. Unlike this paper, they need not, and do not, analyze in detail to what extent a DNN alone contains an explanation. When it comes to the explanatory power of DNNs themselves, Jebeile et al. (2020) are in fact skeptical. In their terms, DNNs have poor representational accuracy because they are “[i]n a nutshell […] statistical techniques that try to reproduce input–output patterns” and are “not designed to capture the processes producing the output variables” (p. 1893). Jebeile et al. are more pessimistic about this criterion than I am because I allow for the possibility that DNNs contain explanations. Only a closer investigation of the network can discover whether they do. In particular, researchers must learn whether the network has become sensitive to variables that figure in explanations. Accordingly, a deeper understanding of the DNNs used is required if they are supposed to provide understanding, precisely as I have argued.

In contrast to Jebeile et al. (2020), Knüsel and Baumberger (2020) study a self-contained machine learning model. Their example is a random forest model, trained on data about forcing (roughly, changes in the net energy flux into the atmosphere) and monthly changes in the Earth’s mean surface temperature. They ran the model once with anthropogenic forcing (which is in fact due to anthropogenic greenhouse gas emission), and once without; the observed increase in temperature was reproduced only in the first case. If the model is representationally accurate, we might use Mill’s method of difference to infer that global warming is at least partially caused by anthropogenic forcing (Mill, 1843/1996, Book III, Ch. VIII). This supports the claim that the model does contain an explanation of global warming.

Knüsel and Baumberger’s ML model is not a DNN, but similar results could be produced using DNNs. Meskhidze (2023) does consider network models: In one of her examples, a network model emulates the results of cosmological computer simulations (Agarwal et al., 2012, 2014). According to Meskhidze, the emulator can account for why the distribution of matter in the Universe has certain features; running it many times can teach us which values of the input parameters (here: the cosmological parameters) yield the observed features of the Universe.

Pietsch (2021, Sec. 4.3.3) provides some background for the method of the two case studies I have just described. In his terms, DNNs (and other ML methods) perform variational induction, a method designed to uncover difference makers and, thus, causes. His argument is that DNNs instantiate traits typical of this kind of induction—for example, they consider positive and negative instances (examples where a phenomenon does or does not occur) and they eliminate irrelevant variables. Pietsch’s account suggests that a DNN trained to predict phenomena under a range of conditions has learned what causal factors are relevant to those phenomena. This, in turn, suggests that DNNs do contain causal explanations—so it is no surprise that scientists can extract causal factors from a DNN.

I agree that, in the case studies from climate science and cosmology, ML models (or DNNs, for our purposes) are used to infer information that is relevant to constructing explanations. Since the examples are best conceptualized using causal explanations, we may adopt causal jargon and say more precisely that a partial cause is identified. This is certainly a sense in which DNNs can be said to increase or provide understanding. However, this sense is different from that used in the simple setting, which focuses on one run of a DNN. In the case studies, the researchers have used or considered several runs as a basis for further inference. The fact that there is typically no understanding in the simple setting is fully compatible with the thought that researchers can use DNNs to gather information that may contribute to an explanation.

Also, the method used in the case studies is not failsafe, for two reasons. First, it presupposes that the model accurately tracks the connection between the variables describing the explanandum phenomenon and whichever variables the researchers have changed. In Meskhidze’s example, it is plausible that this condition is fulfilled; the network has been trained using outputs from computer simulations that accurately represent the underlying physics. We cannot say something like this about all kinds of DNNs, including those that merely reproduce a phenomenon, as in the simple setting.

Second, even if an input variable’s correlation with the explanandum phenomenon is correctly identified using a DNN, it does not follow that an explanatorily relevant factor (e.g., a partial cause) has been detected. This is because Mill’s method of difference only works properly if the homogeneity condition is fulfilled, i.e., if changing an input variable does not affect (other) partial causes of the explanandum phenomenon. This condition may often be fulfilled for big data, if they cover all causally relevant variables (Pietsch, 2021, p. 50). However, data input to DNN models often includes only a few variables, so researchers cannot exclude the possibility that a change in one input variable is, for example, a side effect of a change in a partial cause not covered in the model. Researchers can only infer that the homogeneity condition is fulfilled if they already know a lot about the correct explanation.

A final limitation of the method is that even if a partial cause of a phenomenon has been identified using an ML model, this is a far cry from a full explanation. For instance, in the example from climate science, after applying the method, researchers do not know any new law of nature, or any mechanism that leads from the decisive input variable (anthropogenic forcing) to an increase in temperature. This is not too much of a problem; we know that there are laws or mechanisms that connect anthropogenic forcing and the increase in temperature. Still, if we had only the ML model, we would be far from knowing the explanation. We would, at most, know one causally relevant factor.

All in all, the case studies by Knüsel and Baumberger (2020) and Meskhidze (2023), as well as Pietsch’s argument, show that DNNs can be used to infer explanatorily relevant factors and, in this sense, enhance understanding. These factors may not only relevant to specific token events, but also to types of phenomena found in a number of DNN runs. However, this method is different from the simple setting, and its successful application requires some prior knowledge about the explanation of the explanandum phenomenon; using the method does not give researchers an explanation from scratch. Interestingly, though, its successful application does not require the researchers to have any detailed knowledge about the internal workings of the DNN as such. They need only know that the DNN gets certain correlations right and that no causally relevant factors are omitted.

Another author who is optimistic about the use of ML models for understanding is Sullivan (2022). She thinks that the opacity of current DNNs is not much of an impediment when we try to use the models to understand real-world phenomena. Rather, we should worry about what she calls “link uncertainty,” which Sullivan defines as “a lack of scientiﬁc and empirical evidence supporting the link that connects the model to the target phenomenon” (p. 112). She also tries to show that some machine learning models have got quite far toward understanding. However, as Räz and Beisbart (2022) have stressed, and unlike this paper, she does not consider how-actually explanations, which are crucial for explanatory understanding. Thus, while the ML models Sullivan examines may yield more insights into the target phenomena, enhancing researchers’ objectual understanding, they do not explain the phenomena they predict. Consider, for instance, the Deep Patient Model that Sullivan uses as an example (Miotto et al., 2016). This model is quite successful at predicting diseases using personal data from electronic health records. Accordingly, it is fair to say that the input data contain factors that are either causally relevant or indicative of such partial causes. Still, unless researchers investigate the model more closely, they do not understand why a particular patient has a certain disease. To get closer to such understanding, they should at least know which aspects of the input data were decisive in producing output stating that the patient has the disease, so they should employ methods that aim at this level of interpretability. In short, they have to apply the method Pietsch and others describe.

To summarize, the recent philosophical literature shows that DNNs can be instrumental in obtaining understanding and thus, in some senses, can provide it. First, DNNs can enhance objectual understanding by informing researchers about connections in the target system. Second, they can instantiate some of the virtues that are central to explanatory understanding. Third, and more importantly, they can be used to obtain information that may then enter the construction of an explanation. Still, this is not enough to support the claim that they provide explanatory understanding in the way, for example, theories do (see Räz & Beisbart, 2022, for this distinction). In the case studies from the recent literature where explanatorily relevant information was inferred, it was crucial that several inputs were processed through the trained DNN. Together with further inferences, the skillful use of DNNs helped scientists to gain explanatorily relevant insight (for strategies to improve understanding with DNNs, see Krenn et al., 2022; Pirozelli, 2022).

1. Conclusion

Do DNNs provide scientists with understanding? While some authors deny this, others have argued that DNNs can contribute to scientific understanding. In this manuscript, I have clarified that the answer to the question depends on what is meant by “providing understanding” (or similar expressions). One issue that needs clarification is the kind of understanding under discussion. While DNNs can easily give us new insights about connections between aspects of a target system and thus add to our objectual understanding of it, this does not suffice for explanatory understanding, which is more demanding. Even if we concentrate on the explanatory understanding, as I did in this paper, our question is not entirely clear. To explain the feeling that DNNs don’t provide scientists with understanding, I have first investigated what I have called the “simple setting”: A scientist has run a trained DNN to obtain a prediction or classification, formulated as proposition p. My related question was whether, and under which conditions, this working scientist has explanatory understanding of why p is the case. I have argued that a scientist typically has no understanding of why p. In my discussion, I have also carved out what explanatory understanding in this setting requires: Scientists must know whether a DNN that predicts a phenomenon contains an explanation of it and what this explanation is. It is possible that DNNs contain explanations of various types, but often not even scientists know whether a particular DNN contains an explanation. Further, the scientists must accept the explanation and be justified in this attitude. Finally, they must be able to work with this explanation to make the inferences understanding requires. This is often an additional challenge for scientists because the explanations DNNs may contain can be very complicated.

What does this imply for the relationship between DNNs’ opacity and their use in understanding phenomena? Opacity is the central reason why we cannot tell whether a DNN contains an explanation. What is missing is knowledge of what the DNN does with the input information about the target system. Does the DNN somehow arrive at relevant laws, at a mechanism underlying p, etc.? We cannot know this without improving our higher-level descriptions of what DNNs are doing with the information about the target system. The lack of better descriptions of the work of DNN in terms of variables or processes referring to the target system is the central obstacle to understanding.

This may look like what Sullivan (2022) calls link uncertainty, viz. uncertainty about the relation between the DNN and the target system, but things are not that straightforward. There may be low link uncertainty because there is strong evidence that the model accurately tracks correlations in the target. But this low level of link uncertainty would offer no help regarding explanations. The real problem with explanatory understanding is, first, that we do not know whether the DNN arrives at anything that might qualify as an explanation of the target phenomenon. This issue is independent of link uncertainty. If the DNN does contain a possible explanation, a further question is, second, whether this is a how-actually explanation. This question being open may be considered a part of link uncertainty, but Sullivan’s definition of link uncertainty isn’t entirely clear about this.

Nor is the problem what Humphreys (2009) calls opacity. A computer simulation that is highly opaque, due to its complex computations, is often still known to contain the relevant explanatory laws. The crucial problem is the lack of informative higher-level descriptions of what the model does that describe it in terms of information about the target.

Still, what I have argued about the simple setting is not the whole truth about this paper’s question simply because this setting does not cover every desirable use of DNNs. As many examples from scientific practice show, further investigating target systems with DNNs may be instrumental in improving understanding of the target phenomena. In particular, running DNNs with several inputs can help identify explanatorily relevant factors. This accounts for those voices who have argued that DNNs can provide understanding in some sense.

All in all, networks may dig quite deep by containing explanations or explanatory information, but this is not sufficient to provide humans with explanatory understanding. We humans ourselves have to dig deeper to understand a target system, so the ball is in our court. How far DNNs can be pushed towards giving us understanding depends on what we humans do with them.

References

Agarwal, S., et al. (2012). PkANN – I. Non-linear matter power spectrum interpolation through artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, *424*(2), 1409–1418.

Agarwal, S., et al. (2014). PkANN – II. A non-linear matter power spectrum interpolator developed using artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, *439*(2), 2102–2121.

Angelov, P., & Soares, E. (2020). Towards explainable deep neural networks (xDNN). *Neural Networks*, *130*, 185–194.

Baumberger, C. (2019). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science*, *50*(3), 367–388.

Baumberger, C., Beisbart, C., & Brun, G. (2016). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 1–34). Routledge.

Baumberger, C.; Brun, G. (2016). Dimensions of Objectual Understanding. In S. Grimm, C. Baumberger, & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemology and philosophy of science* (pp. 165–189). Routledge.

Beer, D. (2023, April 7). *Why humans will never understand AI*. Machine Minds – BBC Future. Retrieved July 21, 2023, from <https://www.bbc.com/future/article/20230405-why-ai-is-becoming-impossible-for-humans-to-understand>

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

Beisbart, C., & Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, *17*(6), Article e12830.

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

Bogen, J, & Woodward, J. (1988). Saving the phenomena. Philosophical Review, *97*(3), 303–352.

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

Buijsman, S. (2023). Causal scientific explanations from machine learning. *Synthese*, *202*, Article 202. <https://doi.org/10.1007/s11229-023-04429-3>

Burrell, J. (2016). How the machine “thinks”: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), Article 2053951715622512.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

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

De Regt, H. W. (2020). Understanding, values, and the aims of science. *Philosophy of Science*, *87*(5), 921–932.

De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*, 137–70.

Duede, Eamon (2023). Deep learning opacity in scientific discovery. *Philosophy of Science*, *90*(5), 1089–1099.

Elgin, C. Z. (2009). Is understanding factive? In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 322–330). Oxford University Press.

Epstein, J. M., & Axtell, R. (1996). *Growing artificial societies: Social science from the bottom up*. Brookings Institution Press/MIT Press.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, *71*(1), 5–19.

Frigg, R., and Hartmann, S. (2020). Models in science. In E. N. Zalta (Ed.), Stanford encyclopedia of philosophy (Spring 2020 Ed.). Stanford University. <https://plato.stanford.edu/archives/spr2020/entries/models-science/>

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*, 5742–5751.

Grimm, S. R. (2001). Ernest Sosa, knowledge, and understanding. *Philosophical Studies*, *106*(3), 171–191.

Grimm, S. R. (2021). Understanding. In E. N. Zalta (Ed.), Stanford encyclopedia of philosophy (Summer 2021 Ed.). Stanford University. <https://plato.stanford.edu/archives/sum2021/entries/understanding/>

Grüne-Yanoff, T. (2009). The explanatory potential of artificial societies. *Synthese*, *169*, 539–555. <https://doi.org/10.1007/s11229-008-9429-0>

Hartmann, S. (1999). Models and stories in hadron physics. In M. Morrison & M. Morgan (Eds.), *Models as mediators: Perspectives on natural and social science.* Cambridge University Press (pp. 52–326).

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. The Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175.

Hills, A. (2016). Understanding why*. Noûs*, *50*, 661–688. <https://doi.org/10.1111/nous.12092>

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese,* *169*, 615–626.

Jebeile, J., Lam, V., & Räz, T. (2020). Understanding climate change with statistical downscaling and machine learning. *Synthese*, *199*, 1877–1897.

Karpathy, A. (2019), A Recipe for Training Neural Networks, Andrej Karpathy blog, <https://karpathy.github.io/2019/04/25/recipe/>.

Kaznatcheev, A. (2013, June 5). *Machine learning and prediction without understanding*. Theory, Evolution, and Games Group. Retrieved January 8, 2023, from <https://egtheory.wordpress.com/2013/06/05/prediction-vs-understanding/>

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531.

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

Krenn, M., Pollice, R., Guo, S. Y., et al*.* (2022). On scientific understanding with artificial intelligence. *Nature Reviews Physics*, *4*, 761–769. <https://doi.org/10.1038/s42254-022-00518-3>

Kuorikoski, J. (2011). Simulation and the sense of understanding. In Humphreys, P., & Imbert, C. (Eds.), Models, simulations, and representations. Routledge.

Kvanvig, J. (2003). *The value of knowledge and the pursuit of understanding*. Cambridge University Press.

Kyburg, H. E. (1965). Salmon’s paper. *Philosophy of Science*, *32*(2), 147–151. <http://www.jstor.org/stable/186407>

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., et al. (2023). Learning skillful medium-range global weather forecasting. *Science*, *382*(6677), 1416–1421.

Lawler, I. (2021). Scientific understanding and felicitous legitimate falsehoods. *Synthese*, *198*(7), 6859–6887.

Lawler, I., & Sullivan, E. (2021). Model explanation versus model-induced explanation. *Foundations of Science*, *26*(4), 1049–1074.

Lemos, P., Jeffrey, N., Cranmer, M., Ho, S., & Battaglia, P. (2023). Rediscovering orbital mechanics with machine learning. *Machine Learning: Science and Technology*, *4*(4), Article 045002.

Lenhard, J. (2019). *Calculated Surprises*. Oxford University Press.

Lewis, D. (1986). Causal explanation. In *Philosophical Papers* (Vol. 2). Oxford University Press.

Liu, H., Yin, Q., & Wang, W. Y. (2018). *Towards explainable NLP: A generative explanation framework for text classification*. arXiv. <https://arxiv.org/abs/1811.00196>

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, *67*(1), 1–25. <https://doi.org/10.1086/392759>

Majszak, M., & Jebeile, J. (2023). Expert judgment in climate science: How it is used and how it can be justified. *Studies in History and Philosophy of Science*, *100*(C), 32–38.

Mantzavinos, C. (2016). *Explanatory pluralism*. Cambridge University Press.

Meskhidze, H. (2023). Can machine learning provide understanding? How cosmologists use machine learning to understand observations of the universe. *Erkenntnis*, *88*(5), 1895–1909.

Meyes, R., et al. (2019). Ablation studies to uncover structure of learned representations in artificial neural networks. *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Mill, J. S. (1996). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation* (J. M. Robson, Ed.). Routledge. (Original work published 1843.)

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*(1), 1–10.

Pietsch, W. (2021). *Big data*. Cambridge University Press. <https://doi.org/10.1017/9781108588676>

Pirozelli, P. (2022). Sources of understanding in supervised machine learning models. *Philosophy & Technology*, *35*(23). <https://doi.org/10.1007/s13347-022-00524-4>

Price, I., Sanchez-Gonzalez, A., Alet, F., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*, 84–90. <https://doi.org/10.1038/s41586-024-08252-9>

Räz, T., & Beisbart, C. (2024). The importance of understanding deep learning. *Erkenntnis* 89, 1823–1840. <https://doi.org/10.1007/s10670-022-00605-y>

Salmon, W. C. (1971). Statistical explanation. In W. C. Salmon & R. C. Jeffrey (Eds.), *Statistical explanation & statistical relevance* (pp. 29–87). University of Pittsburgh Press.

Salmon, W. C. (1990). *Four decades of scientific explanation*. University of Minnesota Press (first published 1989).

Schubbach, A. (2021). Judging machines: Philosophical aspects of deep learning. *Synthese*, *198*, 1807–1827. <https://doi.org/10.1007/s11229-019-02167-z>

Scriven, M. (1962). Explanations, predictions, and laws. In H. Feigl & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science: Vol. 3. Scientific explanation, space, and time* (pp. 170–230). University of Minnesota Press.

Sullivan, E. (2022). Understanding from machine learning models. *British Journal for the Philosophy of Science*, *73*(1), 109–133

Tamir, M., & Shech, E. (2023). Machine understanding and deep learning representation. *Synthese*, *201*, Article 51. <https://doi.org/10.1007/s11229-022-03999-y>

Teller, P. (1980). Computer proof. *Journal of Philosophy*, *77*(12), 797–803.

Trout, J. D. (2002), Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*(2), 212–233. <https://doi.org/10.1086/341050>

Tymoczko, T. (1979). The four-color problem and its philosophical significance. *Journal of Philosophy*, *76*(2), 57–83.

Vasudevan, R. K. et al. (2011), Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality. *npj Computational Materials* **7**, 16. <https://doi.org/10.1038/s41524-020-00487-0>.

Weyn, J. A., et al. (2021). Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, *13*(7), Article e2021MS002502.

1. For a closer discussion of the fact that many explanations target at *phenomena*, see Sect. 3.a. [↑](#footnote-ref-1)
2. As long as causality is not merely defined in terms of the networks’ structures. [↑](#footnote-ref-2)
3. The distinction between data and phenomena may be used to construct a quick argument for the claim that DNNs provide no explanatory understanding: A DNN yields data-like output, not a statement about a phenomenon, and thus not a statement about the explanandum phenomenon. Accordingly, the DNN’s inference cannot be reconstructed using an argument that explains why the explanandum phenomenon has occurred. This argument contains the valuable insight that scientists often have to make a non-trivial step to get from a DNN’s output to the explanandum phenomenon. Still, this step, or the argument just described, cannot be the main reason why DNNs do not provide explanatory understanding. And that argument is not, in fact, my main reason. My main point is that the DNN is not known to contain an explanation. The knowledge that is lacking here is different from the knowledge needed to extract the phenomenon from the DNN’s output. [↑](#footnote-ref-3)
4. In realistic examples, this knowledge about the behavior of the DNN will be quite general, thus enabling the scientist to run inferences that are crucial for explanatory understanding of *types* of phenomena in the target system. [↑](#footnote-ref-4)