Interdisciplinary Description of Complex Systems 22(1), 59-70, 2024

DEEP LEARNING IN SCIENCE: IS THERE A REASON FOR (PHILOSOPHICAL) PESSIMISM?

Martin Justin

University of Ljubljana, Faculty of Arts, Department of Philosophy
Ljubljana, Slovenia

DOI: 10.7906/indecs.22.1.3
Regular article

Received: 27 May 2023.
Accepted: 5 February 2024.

ABSTRACT

In this article, I will review existing arguments for and against this philosophical pessimism about using deep learning models in science. Despite the remarkable results achieved by deep learning models networks in various scientific fields, some philosophers worry that because of their opacity, using these systems cannot improve our understanding of the phenomena studied. First, some terminological and conceptual clarification is provided. Then, I present a case for optimism, arguing that using opaque models does not hinder the possibility of gaining new understanding. After that, I present a critique of this argument. Finally, I present a case for pessimism, concluding that there are reasons to be pessimistic about the ability of deep learning models to provide us with new understanding of phenomena, studied by scientists.

KEY WORDS

deep learning, scientific understanding, explanation, black box problem, artificial neural networks

CLASSIFICATION

JEL: O33

* Corresponding author, η: martin1123581321@gmail.com; +386 51 233 505;
Bratovševa Ploščad 22, 1000 Ljubljana, Slovenia.

INTRODUCTION

In the last decade, artificial neural networks (ANNs) using deep learning (DL) achieved some remarkable results when used in for scientific research. AlphaFold a model developed by DeepMind learned to accurately predict possible novel protein structures [1]. DL models are used to successfully predict the presence of breast cancer [2] and the one-, two- and three-year risk of lung cancer [3] from images. DL has proven useful in helping researchers to identify the correlations between the structure and properties of materials in the field of materials science [4]. In neuroscience, such models perform much better than the previously existing ones to model neural single-unit and population responses in higher visual cortical areas [5].

The success of deep learning models is due in large part to their increased complexity [6]. Buckner [7] compares the difference between older, "shallow", ANNs and DL models to the difference between a small team of engineers assembling a car and doing it on a mass-production assembly line. Although the process is in principle the same, the assembly line is exponentially more efficient and reliable. The same goes for deeper neural networks. Bucker writes: "Similar gains in the efficiency and complexity of representational schemes and decision-making policies are afforded by additional depth in neural networks. Specifically, deeper networks can solve certain types of classification and decision problems exponentially more efficiently than shallower networks" [7; p.3].

But increased complexity also means that human users have less understanding of how these models work. As stated in a recent survey of literature on explainable artificial intelligence (XAI): "Though they appear powerful in terms of results and predictions, AI algorithms suffer from opacity," meaning "that it is difficult to get insight into their internal mechanism of work" [8; p.52138]. Given this, some philosophers worry that despite their predictive accuracy, using these systems cannot improve our understanding of the phenomena studied. In other words, they ask whether the fact that we cannot understand highly complex models means that we cannot use them to gain new understanding of the phenomena they predict[1].

In this article, I will review existing arguments for and against this philosophical pessimism about using DL models in science. In what follows, I will first provide some clarification on the terms like artificial intelligence, neural networks, machine learning, the black box problem, explanation, and understanding. In section 3, I will review an argument, presented by Sullivan [9], that opaque ANNs can provide us with understanding. In section 4, I will present some critiques of her argument. In section 5, I will present a positive argument for pessimism about the ability of ANNs to provide us with understanding Finally, I will conclude by stating that the argument for pessimism is convincing, albeit with some qualifications.

SETTING THE SCENE

ARTIFICIAL NEURAL NETWORKS AND DEEP LEARNING

In this section, I will present some terminological and conceptual clarifications that will be important for the discussion in the later parts of the article. Artificial neural networks or ANNs represent one approach in AI research [10]. Unlike AI systems based on deductive logic and probabilistic reasoning, artificial neural networks "are based on formalisms that can be broadly termed 'neurocomputational' " [10]. The idea is that the structure of ANNs was inspired by the structure of organic neural networks, abstracted to the level of a mathematical formalism. ANNs are thus composed of what are called "neurons", which are "simple, unintelligent units that are interconnected by weighted nodes" [11; p.42]. "Neurons" or units are usually layered into a multi-layer net, with each unit being connected to some or all units in the next layer. Each unit also has a so-called activation value. This value is calculated in two steps. First, the input function calculates the weighted sum of all input values, i.e., the values of the weighted nodes connecting the unit to the previous layer. Then, the activation function converts this value into the activation level, i.e., the output, of the unit [10].

Such networks are “almost exclusively used for building learning systems” [10]. In other words, they are used to build systems that “adapt to the environment through learning, which takes place according to a chosen learning rule. Using the learning rule, the system gradually changes the strength of the connections between units” [11; p.45]. Artificial neural networks thus play an important role in the research field of machine learning that “building systems that improve their performance by solving tasks” [10]. It is important to notice here that some machine learning algorithms do not employ ANNs, meaning that machine learning is not a subfield of neurocomputational artificial intelligence. However, one of the biggest recent advances in the design of artificial neural networks has come in the form of the development of deep learning models, i.e., models with increased depth.

BLACK BOX PROBLEM

Let us now turn to the problem of opacity or the so-called black box problem. As already mentioned in the introduction, researchers in the field of XAI have pointed out that “predictive accuracy [of machine learning systems] has often been achieved through increased model complexity” [6; p.1]. This increased complexity, “combined with the fact that vast amounts of data are used to train and develop such complex systems,” has inherently reduced researchers’ ability to “explain the inner workings and mechanisms” of these systems. As a result, “the rationale behind decisions [of these systems] becomes quite hard to understand and, therefore, their predictions hard to interpret”. Therefore, they say that "there is clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions" [6, p. 1]. Adadi and Berrada [8] have reached a similar conclusion. They state that “there are algorithms that are more interpretable than others are, and there is often a trade-off between accuracy and interpretability: the most accurate AI/ML models usually are not very explainable (for example, deep neural nets, boosted trees, random forests, and support vector machines), and the most interpretable models usually are less accurate (for example, linear or logistic regression)” [8; p.52145].

Linardatos, Papastefanopoulos, and Kotsiantis [6] thus distinguish between “black-box” models, which have state-of-the-art performance but are opaque, and “white-box” or “glass-box” models, which are more easily interpretable, but not as powerful. Chirimuuta [12] also specifies which aspects of deep neural networks suffer from opaqueness. She argues that scientists know the activation values of the units, the learning rule, the depth of the network, and the connectivity between the layers. But they do not know exactly how an already trained network arrives at a prediction or classification.

EXPLAINABILITY, INTERPRETABILITY, AND UNDERSTANDING

The problem of opacity has encouraged researchers to search for ways of making ANNs and especially DL models more understandable to human users. But one salient feature of the literature on explainable AI is the imprecise or even interchangeable use of the concepts of explainability, intelligibility, and interpretability [13]. This is also recognized by the researchers themselves. For example, Linardatos, Papastefanopoulos, and Kotsiantis observe that there is “not a concrete mathematical definition of explainability and interpretability” [6; p.2]. Nevertheless, they make a conceptual distinction between these two terms. Interpretability, on the one hand, is understood in connection to the ability of researchers to intuitively understand the relationship between the inputs and outputs of a system. Explainability, on the other hand, is understood in relation to understanding the inner workings of a system. In contrast, Gilpin et al. [14] define explainability as the possibility to provide a satisfactory answer to the "w-questions" regarding the functioning of a system. They also make a difference between two levels of explanation, connected to two different questions scientists can ask about a system. A question about the relationship between inputs and outputs, i.e., "why does this particular input lead to that particular output?" And a question about the internal workings of a system, i.e., "what information does the network contain?"

Definitions of these terms are similarly contested in philosophy. Nevertheless, I think there are at least two distinctions that are important for our discussion. First is a distinction between explainability and interpretability, i.e., between being able to provide an explanation of a given model and being able to understand it. Not everyone accepts this distinction. For example, Beisbart and Räz [15] argue that we should understand interpretability and explainability as synonyms. Nevertheless, there is a long tradition in the philosophy of science of separating the two terms [16] that I think should not be so easily dismissed. Thus, to examine this distinction in a bit more detail, I will present an argument, offered by Erasmus, Brunet, and Fisher [17], that the increased complexity of DL systems affects their understandability or interpretability but not their explainability.

In essence, Erasmus, Brunet, and Fisher [17] argue that (1) the possibility of explaining a given phenomenon P is independent of the complexity of P, while (2) the increased complexity of P hinders our ability to understand it. To argue for (1), they rely on the analysis of explanation which holds that it consists of three elements: (a) the *explanandum*, i.e., what we want to explain, (b) the *explanans*, i.e., with what we are explaining, and (c) the *process of explanation*. Different models of explanations differ in one or more of these elements. Erasmus, Brunet, and Fisher [17] present four such models which feature prominently in the literature. (I) The Deductive Nomological model, in which the explanans includes empirical content plus a law-like preposition, and the process of explanation takes the form of deductive reasoning. (II) The Inductive Statistical model, in which the explanans includes a statistical law about the behavior of the variables, and the process of explanation takes the form of inductive or probabilistic reasoning. (III) The Causal Mechanical model with which scientists aim to show "how the explanandum fits into the causal structure of the world" [17; p.838] and thus involves giving information about the causal process and the causal interaction that leads to the emergence of the explanandum. (IV) The New Mechanist model, in which the explanans includes the entities and their activities that are responsible for the emergence of the explanandum[2].

Then they argue that under all four models of explanation, complexity and explainability are independent[3]. Let us take the Deductive Nomological explanation as an example. According to the above definition, it requires only that the explanans contains a law, and that the process of explanation takes the form of deductive reasoning. It does not matter how complex the two elements are. Thus, an explanation that contains a more complex explanans and requires more complex reasoning may be less desirable, but it is no less an explanation. This also holds, *mutatis mutandis*, for other models of explanation. Consequently, the fact that DL models are increasingly more complex should not, at least principally, affect our ability to explain them.

What about understanding? Erasmus, Brunet, and Fisher [17] point out that authors who study understanding disagree about its exact nature. Nevertheless, they commonly observe that, while an explanation is necessary for understanding, it is not sufficient for it. So, to gain understanding of a phenomenon, some other conditions besides having an explanation must be met. There are several candidates for these additional conditions in the literature, but, as Erasmus et al. argue, they all have in common that they are some "subjective features of the individual who is trying to understand the phenomenon in question" [17; p.848]. One such condition is the criterion of intelligibility. It states that a theory T is intelligible to a scientist in context C if the scientist is able to recognize the qualitatively distinct consequences of T without doing the exact calculations [18]. Given this, it is obvious that the increased complexity of an explanation or a phenomenon makes it less intelligible and thus less understandable. Thus, it can be concluded that complexity affects the ability to understand a phenomenon.

The other distinction that I think is important is a distinction between understanding the models *themselves* and the understanding *provided by* the models. There is an intuitive distinction between having a grasp on how a model works and learning something about the target phenomenon by using the model. This distinction also captures the essence of the disagreement between pessimists and optimists about using DL in science. The optimists, like Sullivan [9], argue that we do not need to understand the models themselves for them to provide us with at least some understanding of the target phenomena. The pessimists, like Chirimuuta [12] or Räz and Beisbart [19], on the other hand, argue that one is a necessary condition for the other.

A CASE FOR OPTIMISM

At face value, there seems to be an intuitive connection between our ability to understand a given model and the possibility that this model provides us with understanding of the target phenomenon. But not all philosophers agree with this line of reasoning. In this section, I will thus present Sullivan's [9] argument that model complexity does not necessarily prohibit understanding.

Her argument can be presented in two parts. First, she argues that having an understanding of a given model is not sufficient for gaining new understanding using this model. Consider, for example, the controversial Schelling's model of segregation. This model is simple and completely transparent. As she describes it: "It is a simulation that consists of a grid with two types of actors, A and B, where both types act on one simple preference—that at least 30 % of their neighbours are the same type. The simulation follows a simple algorithm: if more than 70 % of the actors adjacent to a particular actor are of a different kind, move that actor to the closest unoccupied space" [9]. The equilibrium result of this model is a segregated board which suggests that simple personal preferences, and not systemic discrimination, can cause segregation. But as Sullivan [9] points out, this model does not explain how segregation actually happens, but only how it can possibly happen[4]. Thus, although Schelling's model provides us with an intelligible explanation of the target phenomenon, it does not provide an understanding of it.

What is missing, according to Sullivan [9], is empirical evidence showing that the model accurately represents a real situation. She describes the process of obtaining that evidence as reducing the "link uncertainty", that is, reducing the uncertainty that we can link the behaviour of the model to the behaviour of the target phenomenon. If this is so, having an intelligible explanation is not sufficient for understanding; we need to also consider the empirical evidence connecting the model to the real world. In other words, "the focus should not be unduly placed on how the model works, but instead consider the explanatory question we ask of the model, the role that the algorithm or model plays in the explanation, and the amount, quality, and kind of scientific evidence needed in order to connect the model to the target phenomenon."

In the second part of her argument, Sullivan [9] argues that understanding a given model is also not necessary for gaining any understanding by using this model. The main point here is that "black boxiness" is not a discrete property but that it comes in degrees. She argues that the issue of opacity is an issue of implementation – different parts of the implementation of a given model can be opaque to its builders. But this in principle does not impact the potential of models to provide understanding of phenomena. Consider the following example. Imagine I try to recreate Schelling's simulation on my computer. I might have some knowledge of a programming language, let us say Python, but maybe I am not all that skilled at programming. So, I start building the simulation. I quickly notice that I do not know how to program the function that will tell me whether, for a given actor, at least 30 % of the surrounding actors are of the same type. I go online and, luckily, I find a library containing a function that does exactly that. The documentation says that the function takes a 2-dimensional array (representing the grid with the actors) and the indexes of a given actor as arguments and returns *True* if at least 30 % of the surrounding actors are of the same type as the given actor and *False* if this is not the case. Thus, I do not understand *how* the function works, but I know *what* it does. Happy with my finding, I import the library and implement the function. The simulation works and although one of its main components of it remains opaque to me, I can still plausibly claim that I understand the model as a whole.

These kinds of implementational black boxes are ubiquitous in science. It thus seems reasonable to conclude that they do not hinder understanding. But what about if a given model is opaque at the highest level of implementation, meaning that we only know the inputs and outputs of the system but nothing about what it does and how? Sullivan [9] concedes this kind of opacity would indeed hinder understanding. But she argues that DL models are not opaque in this way. First, as I already mentioned above, we know quite a lot about how DL models work. In addition, Sullivan mentions methods such as salience maps that can help us understand, for example, which features of the input data are the most relevant in the decision-making process of the model. She concludes that "it should be clear from the above discussion that DNNs [deep neural networks] are not black boxed at the highest level either during the modelling process, or in the resulting model" [9].

In the last part of her paper, Sullivan [9] gives some examples of the explanatory value of DL models. She argues that such models can provide us with how-possibly explanations. One example she uses is Deep Patient, a DL model that can accurately predict future health complications based solely on the medical records of patients. Sullivan [9] argues that Deep Patient can provide how-possibly explanations: it shows that it is possible to diagnose patents based only on their medical records. In addition, using salience maps we can determine which parts of the records were especially important for the model's decisions. This way the model could "point to possible correlations that are worthy of future scientific and empirical research" [9]. Given all this, Sullivan [9] argues that we should focus more on testing these correlations and thus reducing the link uncertainty between the model and the real-world phenomena. "The stronger the link, the greater possible understanding the model can provide" [9]. Pessimism about the use of DL models in science is thus unwarranted.

A CASE AGAINST OPTIMISM

In the previous section, I summarized Sullivan's case for optimism about using DL models in science. In this section, I will present some arguments against her case. First, I will present Räz and Beisbart's [19] argument that Sullivan's thesis is tangible only under a very weak notion of understanding. Then, I will use Boge's [20] distinction between two levels of opacity in DL to argue that DL models are not implementational black boxes.

Räz and Beisbart [19] examine three key epistemic insights that Sullivan [9] claims are offered by DL models, and they assess whether these insights indeed enhance understanding. These three insights are: predictive success, how-possibly explanations, and playing a heuristic role in guiding future scientific research. They argue that all of these are necessary conditions for understanding "but that they do not lead to a high degree of explanatory understanding, because they are too far from actual explanations" [19]. As I showed previously, understanding can be analysed as an intelligible explanation. But under Sullivan's [9] account, DL models provide as neither with actual explanation nor are they highly intelligible to us. In the best case, we have a how-possibly explanation and some degree of intelligibility which rests on the possibility to examine the models with post hoc techniques such as salience maps.

In addition, Räz and Beisbart [19] question whether DL models provide us with how-possibly explanations in the same way as more traditional simulations. As we saw, Schelling's model shows one possible way how segregation could happen. Can we say something similar about the Deep Patient model Sullivan mentions? Sullivan writes: "The model can also be used to explain how it is possible to predict schizophrenia (or any of the other seventy-seventy medical problems) through past medical records alone. Simply having a highly predictive model, and knowing the high-level emerging properties of the model, uncovers that it is possible to use a machine learning representation for disease prediction" [9]. But as Räz and Beisbart [19] point out, there seems to be a fundamental disanalogy between the cases. Instead of providing a how-possibly explanation of the target phenomenon, i.e., the relation between schizophrenia (or any of the other problems) and existing medical history, Deep Patient answers "a question about the possibility of predictive modeling itself" [19]. In other words, it answers a methodological rather than substantive question.

Another problematic aspect of Sullivan's argument is her assertion DL models are opaque in essentially the same way as other kinds of models used in science. This ignores one important distinction in the XAI literature mentioned above. Namely, the distinction between answering the question "Why does this particular input lead to that particular output?" in contrast to answering the question "What information does the network contain?" [14].

This distinction is further explicated by Boge [20]. He begins his exposition of the two aspects of opacity by defining opacity. He uses Humphreys's definition of epistemic opacity which states that: "a process is epistemically opaque relative to a cognitive agent X at time *t* just in case X does not know at *t* all of the epistemically relevant elements of the process" [21; p.618]. Boge [20] then distinguishes between two aspects of the opacity of deep neural networks. First, he describes h-opacity. This kind of opacity concerns the workings of a system: a system is h-opaque if it is the process of its operation that is not intelligible to its human users. This is the kind of opacity that results from the complexity of deep neural networks and hinders the understanding of the connection between input and output data. But as Boge [20] notes, this type of opacity is not qualitatively different from, say, the opacity of other complex computational simulations, e.g., climate simulations. In other words, this is what Sullivan calls an implementational black box. But Boge [20] identifies another aspect of opacity that is specific to deep neural networks. He calls is w-opacity. W-opacity concerns the representational content of the system (what was learned). According to Boge, in DL models, not only the process that takes a neural network from an input to an output, but also the properties of the input data that guide this process are opaque.

This distinction is important from the point of view of gaining understanding via DL models. H-opacity only hinders the understanding of the computational model itself, as it prevents researchers from seeing how it gets from input to output data. In contrast to this, w-opacity reduces the potential of deep neural networks to bring new understanding of the target phenomena. Even in the case where promising results would suggest that a DL model "discovered" an important feature of input data that was previously missed by researchers, w-opacity would make this discovery incomprehensible to scientists. It could happen, for example, that the Deep Patient model would discover an important new correlation between some aspect of a given patient's medical history and risk for schizophrenia. But since this model is w opaque, scientists must depend on methods, such as salience maps[5], to try to extract this information. This, together with the above argument about the weak reading of understanding, in my view, defeats Sullivan's argument for optimism about using DL in science.

A CASE FOR PESSIMISM

In In this final section, I will present Chirimuuta's [12] case for pessimism about using DL in science. In her paper, she focuses on using ANNs in computational neuroscience, but I think that with some qualifications, her main points can be generalized to other areas of scientific research.

Chirimuuta defines computational neuroscience as "a tradition of research that builds mathematical models of neurons' response profiles, aiming both at predictive accuracy and at theoretical understanding of the computations performed by classes of neurons" [12; p.771]. It is based on the assumption that information about the external world is "encoded" in the electrical and chemical signals of the neurons. Researchers in the field thus attempt to solve the so-called "decoding problem", i.e., to find a mathematical function that could successfully link neuron spikes to outside information. Specifically, according to Chirimuuta [12], they try to devise a theory of how neurons encode information about the outside world and then write a program, called an encoder, that performs the translation operation between the stimuli and the neural activity.

Thus, as Chirimuuta [12] points out, computational neuroscience pursues two separate epistemic goals. On the one hand, it aims at accurately predicting the relations between neural activity and external stimuli (e.g., to predict how neurons will fire if we show a picture of a square to a primate). On the other hand, it tries to understand how this translation takes place. Chirimuuta [12] thus argues that in the past, when even very simple linear models have proved surprisingly accurate in certain contexts, there has been a convergence between these two goals. However, with the development of deep neural networks, which are much more accurate but w-opaque, these two goals started to diverge.

Chirimuuta [12] presents two examples of such divergence, one from modeling the functioning of the motor cortex and another from modelling the visual perception system. I will limit my presentation to the former, i.e., to her comparison between two studies that tried to model motor cortex activity, Georgopoulos et al. [22] and Sussillo et al. [23]. In both experiments, researchers measured the activity of individual neurons in non-human primates while the primates were performing given tasks. Georgopoulos et al. [22] present an experiment in which a monkey was surrounded by eight buttons, with the ninth button in front of her. In the experiment, first, the button in front of the monkey lit up. After the monkey held it for one second, one of the other eight buttons lit up, and the monkey had to press it with the same hand. Meanwhile, the scientists measured the activity of a population of neurons in her motor cortex and tried to establish a correlation between this activity and the direction of her arm movement. They did this by simply converting the activity of a neuron into a vector in three-dimensional space according to a formula they devised, and then summing the vectors of the individual neuronal cells to obtain one vector that represented the whole neuron population. They found out that the direction of this vector quite closely matched the direction of arm movement. Because of the fairly simple math they used, their model was completely intelligible. In addition, the researchers themselves determined which information about the neural activity is important and should be used to calculate the movement vector. The accuracy achieved by the model can thus be seen as a partial confirmation that these features of neural activity are indeed important for directing arm movement.

The experiment reported by Sussillo et al. [23] is a bit different. They also had non-human primates, this time two, implanted with electrodes that measured the activity of individual neurons in their motor cortex. But the monkeys did not press buttons; rather, they had to move a cursor on a screen from a central position to a marked position in one of the corners of the screen. Each monkey performed three series of experiments. First, they moved the cursor by moving their hand. Then, they moved the cursor using a brain-machine interface (BMI) that used an encoder, based on a mathematical model, similar to the one described in the previous example. In the last series, they used a BMI that encoded information using a trained neural network. Each monkey performed each of the three experiments hundreds of times. The researchers found that using this ANN-based encoder significantly improved the monkeys' performance vis-à-vis the older model. This suggests that the BMI with an ANN was more successful in translating between neuronal activation and information about the outside world. We can thus assume that the ANN either has approximated the mathematical function linking neuronal activation and external stimuli more accurately or it has "discovered" new features of the input data that play an important role in the translation. But the ANN used was both h-opaque and w-opaque so despite its improved performance, it did not provide scientists with a better understanding of how the motor cortex works.

It should now be clear what Chirimuuta [12] is getting at when she says that the use of complex ANNs creates a divergence between the goals of predictive accuracy and understanding of neurological processes. It should be noted again that researchers can use post hoc methods such as salience maps to extract at least some information from the models [9, 24, 25]. Nevertheless, there is a clear sense in which the opacity of DL models used affects the ability of scientists to gain new understanding black boxes.

CONCLUSION

Some philosophers worry that because of their opacity, deep learning systems cannot improve our understanding of the phenomena studied. In this article, I tried to establish whether there are good reasons for this philosophical pessimism about using deep learning models in scientific research. First, I clarified the main concepts used in the article. Next, I presented a case for optimism about using DL in science, as found in Sullivan [9]. After that, I presented a critique of her arguments, relying mostly on Räz and Beibart's [19] and Boge's [20] discussion of the opacity of DL models. Finally, I summarized a case for pessimism, as presented in Chirmuuta [12]. The article concludes that despite the predictive success of DL models, there are reasons to be pessimistic about the ability of DL models to provide us with new understanding of phenomena, studied by scientists.

REMARKS

[1] There is also a huge body of literature on the ethical issues of using opaque technologies in decision-making processes. These issues are especially pronounced in domains where such decisions affect individuals' well-being, e.g., in medicine or the justice system. For a review of this literature, cf. Mittelstadt et al. [26]; Jobin, Ienca, and Vayena [27]. I will not consider this problem here.

[2] Woodward and Ross [28] present a slightly different typology. In particular, they add Salmon's statistical relevance model and pragmatic models of explanation.

[3] Prasetya [29] points out that Erasmus, Brunet, and Fisher [17] left out one important model of explanation, i.e., the unificationist model of explanation, which consists in providing a pattern of inference that can be used to describe and thus explain many different phenomena. Prasetya [29] argues, pace the general thesis in Erasmus, Brunet, and Fisher [17], that this kind of explanation is sensitive to the complexity of both the explanandum and the explanans. In a response, Erasmus and Brunet argue that the unificationist model of explanation is dissimilar to the other four models in that it is "about whether a theory is 'explanatory,' while the other familiar accounts of explanation are about whether a given act/argument is an explanation" [30; p.42]. Since in the original paper, they admit that a more complex

explanation might be less desirable or less "explanatory", Prasetya's [29] point about the unificationist model does not defeat their general thesis.

[4] For a more detailed discussion on the distinction between how-possibly and how-actually explanations, cf. Reutlinger, Hangleiter, and Hartmann [31].

[5] It is questionable how reliable such methods actually are. Räz and Beibart for example write: "Saliency maps can indeed be useful. However, there is no guarantee that if a saliency map looks fine, the model is fine. Saliency maps are heuristic tools; they do not provide general understanding of a model" [19]. This is supported by literature in XAI: Linardatos et al. for example write that "one of the issues with saliency maps is that concepts in an image, such as the 'human' concept or the 'animal' concept, cannot be expressed as pixels and are not in the input features either and therefore cannot be captured by saliency maps" [6; p.9].

ACKNOWLEDGMENTS

A shorter version of this article has been published in the "Proceedings of the 25th International Multiconference Information Society, Volume B: Cognitive Science". I am grateful to the participants of the conference for their comments and suggestions on that version of the paper. I would also like to thank Olga Markič for her help and encouragement, and Nejc Hirci for his help with more technical aspects of the literature.

REFERENCES

[1] Jumper, J., et al.: *Highly accurate protein structure prediction with AlphaFold*.
Nature **596**(7873), 583-589, 2021,
http://dx.doi.org/10.1038/s41586-021-03819-2,

[2] Wu, N., et al.: *Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening.*
IEEE Transactions on Medical Imaging **39**(4), 1184-1194, 2020,
http://dx.doi.org/10.1109/TMI.2019.2945514,

[3] Huang, P., et al.: *Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method*.
Lancet Digital Health **1**(7), 353-362, 2019,
http://dx.doi.org/10.1016/S2589-7500(19)30159-1,

[4] Ge, M.; Su, F.; Zhao, Z. and Su, D: *Deep learning analysis on microscopic imaging in materials science*.
Materials Today Nano **11**, No. 100087, 2020,
http://dx.doi.org/10.1016/j.mtnano.2020.100087,

[5] Yamins, D.L.K. and DiCarlo, J.J.: *Using goal-driven deep learning models to understand sensory cortex*.
Nature Neuroscience **19**(3), 356-365, 2016,
http://dx.doi.org/10.1038/nn.4244,

[6] Linardatos, P.; Papastefanopoulos, V. and Kotsiantis, S.: *Explainable AI: A Review of Machine Learning Interpretability Methods*.
Entropy **23**(1), No. 18, 2020,
http://dx.doi.org/10.3390/e23010018,

[7] Buckner, C.: *Deep learning: A philosophical introduction*.
Philosophical Compass **14**(10), 2019,
http://dx.doi.org/10.1111/phc3.12625,

[8] Adadi, A. and Berrada, M.: *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).*
IEEE Access **6**, 52138-52160, 2018,
http://dx.doi.org/10.1109/ACCESS.2018.2870052,

[9] Sullivan, E.: *Understanding from Machine Learning Models*.
British Journal for the Philosophy of Science **73**(1), 109-133, 2022,
http://dx.doi.org/10.1093/bjps/axz035,

[10] Bringsjord, S. and Govindarajulu, N.S.: *Artificial Intelligence*.
Stanford Encyclopedia of Philosophy, 2020,
https://plato.stanford.edu/entries/artificial-intelligence, accessed 28th January 2022,

[11] Markič, O.: *Cognitive Science: Philosophical Questions*. In Slovenian.
Aristej, Maribor, 2011,

[12] Chirimuuta, M.: *Prediction versus understanding in computationally enhanced neuroscience*.
Synthese **199**(1-2), 767-790, 2021,
http://dx.doi.org/10.1007/s11229-020-02713-0,

[13] Krishnan, M.: *Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning*.
Philosophy & Technology **33**(3), 487-502, 2020,
http://dx.doi.org/10.1007/s13347-019-00372-9,

[14] Gilpin, L.H., et al.: *Explaining Explanations: An Overview of Interpretability of Machine Learning*.
In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp.80-89, 2018,
http://dx.doi.org/10.1109/DSAA.2018.00018,

[15] Beisbart, C. and Räz, T.: *Philosophy of science at sea: Clarifying the interpretability of machine learning*.
Philosophy Compass **17**(6), No. e12830, 2022,
http://dx.doi.org/10.1111/phc3.12830,

[16] Grimm, S.: *Understanding*.
The Stanford Encyclopedia of Philosophy, 2021,
https://plato.stanford.edu/archives/sum2021/entries/understanding, accessed 10th February 2023,

[17] Erasmus, A.; Brunet, T.D.P. and Fisher, E.: *What is Interpretability?*
Philosophy & Technology **34**(4), 833–862, 2021,
http://dx.doi.org/10.1007/s13347-020-00435-2,

[18] de Regt, H.W.: *Understanding and scientific explanation*.
In: de Regt, H.W.; Leonelli, S. and Eigner, K., eds: *Scientific Understanding: Philosophical Perspectives*. University of Pittsburgh Press, 2008,

[19] Räz, T. and Beisbart, C.: *The Importance of Understanding Deep Learning*.
Erkenntnis, 2022,
http://dx.doi.org/10.1007/s10670-022-00605-y,

[20] Boge, F.J.: *Two Dimensions of Opacity and the Deep Learning Predicament*.
Minds & Machines **32**, 43-75, 2022,
http://dx.doi.org/10.1007/s11023-021-09569-4,

[21] Humphreys, P.: *The philosophical novelty of computer simulation methods*.
Synthese **169**(3), 615-626, 2009,
http://dx.doi.org/10.1007/s11229-008-9435-2,

[22] Georgopoulos, A.P.; Schwartz, A.B. and Kettner, R.E.: *Neuronal Population Coding of Movement Direction*.
Science **233**(4771), 1416-1419, 1986,
http://dx.doi.org/10.1126/science.3749885,

[23] Sussillo, D., et al.: *A recurrent neural network for closed-loop intracortical brain–machine interface decoders*.
Journal of Neural Engineering **9**(2), No. 026027, 2012,
http://dx.doi.org/10.1088/1741-2560/9/2/026027,

[24] Pirozelli, P.: *Sources of Understanding in Supervised Machine Learning Models*.
Philosophy & Technology **35**(2), No. 23, 2022,
http://dx.doi.org/10.1007/s13347-022-00524-4,

[25] Duede, E.: *Deep Learning Opacity in Scientific Discovery*. preprint arXiv:2206.00520v2 [cs.AI], http://dx.doi.org/10.48550/arXiv.2206.00520,

[26] Mittelstadt, B.D., et al: *The ethics of algorithms: Mapping the debate*. Big Data & Society **3**(2), 2016, http://dx.doi.org/10.1177/2053951716679679,

[27] Jobin, A.; Ienca, M. and Vayena, E.: *The global landscape of AI ethics guidelines*. Nature Machine Intelligence **1**(9), 389-399, 2019, http://dx.doi.org/10.1038/s42256-019-0088-2,

[28] Woodward, J. and Ross, L.: *Scientific Explanation.* Stanford Encyclopedia of Philosophy, 2021, https://plato.stanford.edu/entries/scientific-explanation, accessed 29th January 2022,

[29] Prasetya, Y.: *ANNs and Unifying Explanations: Reply to Erasmus, Brunet, and Fisher*. Philosophy & Technology **35**(2), No. 43, 2022, http://dx.doi.org/10.1007/s13347-022-00540-4,

[30] Erasmus, A. and Brunet, T.D.P.: *Interpretability and Unification*. Philosophy & Technology **35**(2), No. 42, 2022, http://dx.doi.org/10.1007/s13347-022-00537-z,

[31] Reutlinger, A.; Hangleiter, D. and Hartmann, S.: *Understanding (with) Toy Model.* British Journal for the Philosophy of Science **69**(4), 1069-1099, 2018, http://dx.doi.org/10.1093/bjps/axx005.