Epistemic implications of machine learning models in science

Buijsman, Stefan; Durán, Juan M.

DOI
10.4324/9781003205647-39

Publication date
2024

Document Version
Final published version

Published in
The Routledge Handbook of Philosophy of Scientific Modeling

33

EPISTEMIC IMPLICATIONS OF MACHINE LEARNING MODELS IN SCIENCE

Stefan Buijsman and Juan M. Durán

1. Introduction

Machine learning models are quickly gaining ground in scientific practice. A particular story of success is the use of the deep learning model AlphaFold 2 to predict protein folding (Jumper et al. 2021), but examples abound. There is, for example, the usage of deep learning in climate models (Rasp et al. 2018), astronomy (Agarwal et al. 2012), and materials science (Schmidt et al. 2019). Furthermore, a wide range of deep learning models are used in computational neuroscience (e.g., Zhuang et al. 2021; Güçlü and van Gerven 2017). This increased usage of machine learning techniques in scientific research raises important philosophical questions regarding the epistemic implications of such tools. Most prominently, the issue is that many machine learning models fail to represent a target system with a set of equations, as is the case in other types of (process-based) models. To see this, consider the workings of deep learning models such as random forest models. These models, a type of neural network, consist of a large number of artificial neurons that have a (standardly non-linear) activation function determining the output value of the neuron based on the input values. These artificial neurons are then ordered into (a large number of) layers, with connections from neurons in one layer to neurons in the next layer. It is those connections that matter, as the weights on them—how much the output of a neuron counts toward the input of the next neuron—are adjusted based on training data. Typically, a machine learning model has millions of weights, and the largest neural network models have trillions of such weights that are adjusted in training.

A number of differences from traditional theoretical models and modeling have already become apparent from this very brief description of machine learning models. First and foremost, there are no (explicit) representations of physical quantities in such models. This differentiates machine learning models from other statistical models, where regression based on data may be used, but representations of physical quantities are still present in the model. Furthermore, the adjustment of the weights in machine learning models happens automatically, based on a training set. There are too many such weights to monitor this process directly, nor can the final model be easily inspected to understand its exact functioning. It follows that it is incredibly difficult to tell which patterns the model uses to arrive

DOI: 10.4324/9781003205647-39

at predictions. As a result, machine learning models have a high degree of epistemic opacity, defined as (see also Durán and Formanek 2018; Beisbart 2021):

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

These two differences, and the increased epistemic opacity that results, raise the philosophical question: what is the scientific value of using machine learning models? The answer depends somewhat on the scientific field. In the case of neuroscience, for instance, artificial neurons might be seen as (idealized) representations of physical neurons. Thus, neural networks can be seen to yield scientific understanding in these contexts and can give way to new specifications of functionalism in the philosophy of mind (Section 2). In the other sciences, their status is much more contentious. The statistical nature of machine learning is a more serious concern here, as is the opacity of models and the absence of representations. Can machine learning models yield scientific explanations and, possibly via those explanations, understanding? Are we justified in believing their predictions? As we will see (Section 3), views range from pessimistic, seeing machine learning models as substantially different from other kinds of models, to more optimistic, where epistemic opacity is not necessarily an issue and machine learning models are on the same scale of explainability as other models used in science.

Does this entail new ways of doing science and, as such, novel issues for the philosophy of science? Some answers to these questions are found in the literature on computer simulations. Although one can find some early skepticism about the *scientific* novelty of computer simulations (e.g., Teichroew and Lubin 1966, 724), the general feeling is that computer-based methodologies extend the class of tractable mathematics and representation, thus broadening the range of modeling phenomena (Frigg and Reiss 2009). Fewer agreements are found, however, on the *philosophical* novelty of computer-based research. Famously, Frigg and Reiss (2009) club together four skeptical arguments against "a new metaphysics, epistemology, semantics, and methodology" (595) for the philosophy of science. Humphreys, however, alerts us that an anthropocentric epistemology is no longer viable and that we are required to understand and evaluate the world through "computationally based scientific methods that transcend our own abilities" (Humphreys 2009, 617), as opposed to representations tailored to human cognitive capacities. Within a non-anthropocentric epistemology emerge diverse philosophical issues that, according to Humphreys, have not been addressed by a more familiar philosophy of science. Perhaps the most famous of all is the problem of *epistemic opacity* mentioned earlier. Having said that, the epistemic and methodological implications of using machine learning models are still heavily debated. However, their successful use in the sciences shows that they certainly have a role to play.

2. Neural networks and neuroscience

The application of machine learning in neuroscience is a special case. As opposed to other sciences, neural networks (but no other machine learning techniques) can be argued to contain explicit representations in neuroscience. Artificial neurons represent actual neurons, the weights in neural networks represent the strength of connections between neurons, and so on. There are, of course, a number of differences between neural networks and actual

neurons, as López-Rubio (2018) enumerates: backpropagation is unrealistic for the brain, artificial neurons are far simpler than biological neurons, the brain isn't structured as neatly as neural networks are, activation functions differ, and so on. Despite such differences, López-Rubio considers it plausible that neural networks are representative of the brain, according to a similarity view of model representation:

> [f]rom the current state of research, it is likely that the similarities among biological and artificial features extend from the highest level of description, i.e., the overall inputs and outputs, to a certain intermediate level of description, while the lowest levels such as the electrical signals in the biological synapses do not match well with their artificial counterparts.
>
> *(681)*

In virtue of this similarity between neural networks and the brain, López-Rubio holds that we can formulate an updated version of computationalism he terms neural computational functionalism:

Neural computational functionalism (NCF): the mind is the set of synaptic weights of the brain.

> This is to be interpreted in the sense that: (a) the brain stores synaptic weights in its neural structures, (b) some of those neural structures are organized in a hierarchy of layers, (c) those synaptic weights determine the computation of significant features of progressively higher level as we traverse the neural hierarchy, (d) those features ultimately determine behavior.
>
> *(682–683)*

Neural networks can then clearly function as models of the brain in much the same way that other types of models work. In line with that idea, neural network models would be able to offer explanations of the functioning of the brain. Piccinini makes a concrete suggestion of what such explanations would look like:

> An *explanation by synaptic weights* of a capacity C possessed by a (biological or artificial) neural system S is a set of weights W for C such that S possesses C because S operates according to its stored weights W.
>
> *(Piccinini 2010, 277)*

Neural networks can then clearly offer explanations of the functioning of the brain in this proposal. Such ideas are more widespread, as Miłkowski (2013) and Stinson (2018) similarly argue that neural network models can offer (mechanistic) explanations of the brain. Buckner (2018) even argues that the functioning of (convolutional) neural networks gives an important insight into the way the brain handles concepts. They illustrate a process he calls *transformational abstraction*, where complexity is reduced by iterative transformations into simplified (abstract) representations. This type of abstraction, which occurs in neural networks in order to detect later layers, e.g., the presence of a chair or shovel, is in Buckner's (2018) eyes, also a fitting solution to the question of how humans manage to acquire such concepts from experience. Can these neural networks then function as mechanistic explanations for the visual cortex of our brains?

There are a number of problems laid out by Buckner (2018), such as the fact that neural networks are prone to adversarial examples: Small changes to the input image can cause the model to yield a wildly different output classification. An image of a panda in which a few select pixels are changed might be classified as showing a gibbon, for example. Such adversarial examples are hard to eliminate in neural networks, and yet our brains are clearly not susceptible to them. Buckner (2018, 5367) does not see this as too problematic and argues instead that neural networks are best seen as mechanism sketches (Piccinini and Craver 2011) or as generic mechanisms of the kind Stinson (2018) suggests. Taking the limitations of neural networks into consideration, they still have an explanatory role to play: "DCNNs show that the generic kind of neural mechanism found in mammalian perceptual cortex can learn and deploy abstract category representations using only domain-general mechanisms—vindicating a key theme of empiricism" (Buckner 2018, 5369). Yet, at the same time, neural networks are far worse than we are at generalizing (see Section 3.1) and make very different mistakes in image classification and other tasks than humans. Such substantial differences call for caution when using neural networks as models of, e.g., human concept formation. Further attention to these functional differences is needed before we can see neural networks as explaining our actual higher-level cognitive functions.

Machine learning models, in conclusion, might provide an idealized representation of biological neurons and synapses, and neural networks can act as (mechanistic) explanations of their functioning on an appropriate level of abstraction. There is more work to do on the exact nature of these representations and idealizations and the effect this has on the conclusions that can be drawn from the models. Can neural networks explain higher-level cognitive functions, or do they only provide how-possibly explanations? Does their limited generalizability imply a limit to their role in how-actually explanations? Or will neural networks become one of the dominant modeling tools for neuroscience? This requires further reflection, but there is little doubt that neural networks have an explanatory role to play. That is quite different when applications of machine learning models are considered in other sciences. We, therefore, turn to those other applications now.

3. Machine learning in the other sciences

As mentioned in the introduction, machine learning models present us with difficulties in the other sciences, as they do not contain explicit representations of the physical quantities involved and are epistemically opaque. That is, we typically do not understand why a machine learning model yields a particular prediction as opposed to a different one. Consequently, it is tempting to hold that such models do not yield (scientific) explanations for the phenomena they are trained to predict. Srećković et al. (2022), for example, argue that machine learning complicates the obtaining of two types of explanations: Process explanations and phenomenon explanations. This is uncontroversial for process explanations, which would be explanations of the process that led to a specific model prediction. Machine learning models are typically too complex to survey, and it is a serious challenge to obtain explanations for their outputs. This is widely studied under the name explainable AI (XAI; see Das and Rad 2020 for a review) and is considered an ethical issue for the application of AI.

The more crucial question for the use of machine learning methods in science is whether this also means that explanations of the scientific phenomena that are predicted are not

forthcoming. Srećković et al. (2022, 6) consider such explanations to be unforthcoming due to the lack of causal relations underlying the model predictions: the problem is "the associativity of the method, which involves searching solely for correlations between the features in the data without a theoretical back-up to provide causal relationships, traditionally considered crucial for explanations." The lack of process explanations exacerbates this issue, as it obscures the correlations used by the model to make predictions. These underlying correlations, as a result, cannot be extracted from the model, and so no experiments can be designed to find causal relations. In short, machine learning models, as Srećković et al. (2022) argue, cannot be used to arrive at causal relations linking inputs to outputs, and so do not yield causal explanations. They can be used for (highly accurate) predictions, but not for understanding. However, it is not even clear that predictions of machine learning models will have a similar, or better, epistemic status as those of process-based models. Thus, before diving deeper into the question of explanations, we turn first to the predictions of machine learning models.

3.1 Epistemic status of machine learning predictions

Machine learning models are usually associated with high accuracy. In the case study that Kawamleh (2021) looks at, the model predictions for parametrizations in climate models are reported to be of high accuracy (Rasp et al. 2018). Despite this success on the test set with which the model was evaluated, Kawamleh (2021) argues that the machine learning model fails to generalize to new situations. This limited generalizability of machine learning models is a known problem, as machine learning models often perform badly when presented with input that is (in our view slightly) different than that present in the test set. For example, object recognition systems become highly inaccurate when objects are presented in unusual locations (Rosenfeld et al. 2018), or when they are rotated into an unusual pose (Alcorn et al. 2019). A similar situation occurs in the machine learning model that predicts parameters for climate models. These are trained on input-output pairs generated by a physical model (that is much more computationally intensive to use for long-term climate change modeling). If the situation deviates too much from these training pairs, which one might expect when modeling climate change, then the machine learning model loses its accuracy. As Rasp et al. (2018, 9687) state, "the neural network cannot handle temperatures that exceed the ones seen during training," in this case, an increase of sea-surface temperatures of more than 4 Kelvin. They blame this on overfitting, but as Kawamleh (2021) shows, no machine learning models have managed to generalize on this task to date (and as pointed out above, it is, in fact, a common feature of such models).

Where does this lack of generalizability come from? Kawamleh (2021) blames the lack of representations of physical processes:

> Traditional and cloud resolving parameterizations represented processes directly or indirectly and this process representation has added an irreducible value for the reliability of model predictions because it provides (a) physical/empirical constraints and (b) facilitates forms of model development and evaluation which guard against overfitting.
>
> *(1019)*

Machine learning models do not have this protection against overfitting and instead rely purely on correlations present in the data set generated from running the process-based model on a chosen set of training cases. The upshot is that:

> the trained NNP [machine learning model] fails to learn convection and generalize beyond its training data because it fails to represent the causal convective processes which relate the climate variables of interest. […] The very *representation* of processes adds significant and irreplaceable value for the reliability of climate model *predictions*.
>
> *(Kawamleh 2021, 1019, emphasis in original)*

This matches with explanations of the incredible performance of machine learning models in protein folding, where AlphaFold 2 is largely considered to have 'solved protein folding' because it gives accurate predictions of the folding for almost all protein specifications. Note, however, that "[t]he key to why AF2 works is the fact the library of single domain protein structures is essentially complete" (Skolnick et al. 2021, 4827). It is this lack of outliers compared to the training set that has led to a uniformly strong performance. If it were not for that completeness, there would likely be the same issues with generalizability (and indeed, issues do occur when more than one fold is possible). For:

> AlphaFold has not learned from ligands and is actually not aware of the actual energy minima that are essential for folding in real life. In reality, AlphaFold has not solved the folding problem as it would occur in solution or in a cell, but it has provided a practical solution: It has learned the results of folding at the amino acid residue contact level and can, therefore accurately predict a single-chain hemoglobin fold that would never exist on its own or in the absence of the heme cofactor in nature.
>
> *(Perrakis and Sixma 2021, 2–3)*

So, does this issue with the generalizability of machine learning models affect the epistemic status of their predictions? It need not, depending on one's views of justification from machine learning models. We only give a brief overview of the options here. These range from more liberal views, such as that of Beisbart (2017), who holds that one is justified to believe the predictions of a computer simulation (here generalized a bit to machine learning) if one is justified to believe that the computer program works as intended. Verification of this will be very difficult for machine learning models, however, due to their epistemic opacity, so when are we justified to believe that the program works as intended? One can also wonder which intentions are relevant, as intending that the model predicts the phenomenon accurately for a test set is easy to verify, but too limited to be justified in believing its results generally speaking.

Durán and Formanek (2018) are more detailed matters about justification, though from a more externalist standpoint. They hold that one is justified to believe the output of a computer simulation if the model is sufficiently reliable, in their account of *Computational Reliabilism* (which can be generalized to machine learning models):

> (CR) if S's believing p at t results from m, then S's belief in p at t is justified.
>
> where S is a cognitive agent, p is any truth-valued proposition related to the results of a computer simulation, t is any given time, and m is a reliable computer simulation.
>
> *(Durán and Formanek 2018, 654)*

Reliability here is to be understood as more than simply that the model produces correct predictions sufficiently often. Instead, it is a more complex notion where the reliability of a model can be supported by reliability indicators such as verification and validation methods, robustness analyses, a history of (un)successful implementations, and expert knowledge. The account does not tell how these factors fit together, and thus, how to determine when a model is reliable and for what range of cases (e.g., only temperature variations under 4 Kelvin). Such details would need to be delivered by applying computational reliabilism to specific cases.

Finally, Symons and Alvarado (2019) take this idea somewhat further, holding that justifications for the results of computer simulations (i.e., machine learning model predictions) in scientific contexts come with high demands. They are, consequently, fairly pessimistic about machine learning models, as they argue that "trust in simulations should be grounded in empirical evidence, good engineering practice, and established theoretical principles. Without these constraints, computer simulation risks becoming little more than unmoored speculation" (Symons and Alvarado 2019, 57–58). Such grounding is difficult, though what it exactly entails is left unclear. Still, Kawamleh (2021) can be read as an argument that scientific grounding is lacking for those machine learning models, so justified beliefs might be hard to come by. The epistemic status of machine learning predictions is thus a matter of active debate, and there is a need for more specific accounts that can adjudicate specific cases. The lack of representations in these models presents a problem for their generalizability and grounding in established theoretical principles. That, in turn, affects the epistemic status of their predictions. Does it also rule out any hope for scientific explanations and understanding?

3.2 Explanations from machine learning models?

The statistical nature of machine learning models, combined with the opacity of the precise correlations they rely on, are for a number of philosophers good reason to be skeptical of their explanatory prospects. We have already discussed the arguments of Srećković et al. (2022), but López-Rubio and Ratti (2021) make a similar point in the context of molecular biology. They focus on the prospect of mechanistic explanations, the standard account for molecular biology, resulting from machine learning models. They, too, are skeptical that such explanations can be obtained: " If you do molecular biology with machine learning techniques, and if you want to have the best machine learning performances, then you cannot even in principle elaborate fully-fledged mechanistic explanations" (López-Rubio and Ratti 2021, 3152). Not because of technological limitations, but because "the more the size of the model increases, the less the human mind is able to organize the model's components into a causal narrative, which forms the backbone of any mechanistic description with explanatory force" (López-Rubio and Ratti 2021, 3152). As machine learning models rely on a vast number of parameters to achieve high accuracy, the argument goes, that they hinder the formulation of a causal narrative and, thus, of a mechanistic explanation. Here it is the associativity, i.e., the lack of a clear causal link between inputs and model predictions, in addition to the complexity that hinders understanding.

Yet other philosophers do not consider it a given that there are no scientific explanations to extract from machine learning models (primarily seen as involving causal relations, though, importantly, not all philosophical accounts of explanation give a central role to causation). They hold that, at least in some cases, it is possible to acquire these kinds of explanations.

Sullivan (2022) started this line of thought, defending that machine learning models can yield understanding despite their epistemic opacity. She holds that the implementation of a model is often irrelevant to the explanations that can be generated from that model and gives the example of Schelling's model, used to study the causes of segregation. This model holds that a person will move if more than 70% of her neighbors belong to a different group than she does. In a situation with two groups present, this simple rule ultimately leads to a segregated situation, irrespective of the starting situation. How that model is implemented, however, whether on a checkers board (as originally the case) or on a computer, is irrelevant for extracting scientific explanations. What matters for us to obtain explanations of segregation in the real world is whether the model shows a process that actually occurs. In other words, what matters is whether people in real life tend to move when they belong to the minority in a specific neighborhood. If the model links to such a real-world process, then it can provide explanations. If it does not, then it fails to yield scientific explanations. The real problem, according to Sullivan, then, is what she calls *link uncertainty*, where "link uncertainty constitutes a lack of scientific and empirical evidence supporting the link connecting the model to the target phenomenon" (Sullivan 2022, 21). Note, however, that the explanation resulting from the model here also crucially relies on us knowing the process implemented by the model: that people move when 70% of their neighbors are in a different group is built-in in the model (Räz and Beisbart 2022). However, as discussed in the context of epistemic opacity, the knowledge of the implemented process is difficult to obtain from machine learning models. Sullivan (2022), however, is optimistic that, in some simple cases, one can still know enough about the implemented process and reduce the link uncertainty sufficiently to obtain explanations from machine learning models.

Sullivan argues that this is the case for a skin lesion classifier, where a machine learning model classifies moles based on their visual appearance. As there is a strong scientific basis for a link between visual appearance and the type of mole it is (e.g., whether it is a kind of cancer or requires a biopsy), the reasoning goes that the link uncertainty, therefore, is low. The model also receives the input information that is scientifically known to be relevant to the decision, and thus, correlations found based on that information are of interest. Perhaps they do not correspond to causal relations, but Sullivan maintains that such (new) correlations "can further understanding, especially once these newly discovered patterns undergo further investigation" (2019, 24). While she does not discuss how the correlations the machine learning model uses would be identified, deal with the worry that they may be too complex, or how link uncertainty is reduced, the idea that available scientific background information can make machine learning models explanatory has been picked up and developed in further detail by others.

Knüsel and Baumberger (2020) do so in the context of climate change modeling - not for parametrizations, but for models that try to determine if the rise in average temperature is due to human actions. In such a case, they consider it possible for machine learning models to provide understanding. The condition here is that:

> for data-driven models to be useful for understanding phenomena, researchers should be in a position to argue from the coherence of the model with background knowledge to its representational accuracy. This can for example be achieved if important bivariate relationships are known. This sort of reasoning provides exactly the kind of evidence that reduces the link uncertainty discussed by Sullivan
>
> *(Knüsel and Baumberger 2020, 47)*

How does this background knowledge help in modeling historical changes in temperature? First of all, it is a setting where we can approximate the situation quite well using the energy-balance model, consisting of a single differential equation. It coheres with background knowledge, has decent empirical accuracy, is robust, and is easily graspable (as it is only a single differential equation). As such, it can be used to show that human actions are the cause of the temperature rise, as that rise only comes out of the model if the effects of human actions are taken into account. Filter them out, and the average temperature predicted by the model remains stable. We can then explain why the average temperature has risen (and why human actions are the culprit).

Knüsel and Baumberger (2020) then compare this process-based model to a machine learning model making the same predictions. This machine learning model shows the same difference whether human actions are included or not and has similar empirical accuracy. They argue that it is robust because outputs are similar to the process-based model, in that it is coherent with background knowledge and because the outputs are consistent with the known physical laws (though recall Kawamleh (2021) that robustness and coherence are more complicated), and that manipulating the model and studying the feature importance makes it somewhat graspable. Therefore, they hold that this machine learning model can also be used to explain why the average temperature has risen in the last hundred years. Machine learning models may do worse on all these scales of explanation except for empirical accuracy, but they can still do well enough in some cases to provide explanations. The argument, however, focuses on whether certain input values (human factors) are relevant to the outcome. More interesting, and problematic, given the associativity and opacity of machine learning models, is *why* human actions cause a rise in temperature. Knüsel and Baumberger (2020) do not discuss that question. In addition, it is unclear if the link uncertainty can be reduced sufficiently without a transparent process-based model being available. Only if that is possible would machine learning models add new explanations.

A similar shortcoming can be seen in the work of Jebeile et al. (2021), who look at yet another type of machine learning in climate modeling to argue that said models are on a continuous scale along with other types of models. They argue that their empirical accuracy is often better, but they do worse on intelligibility, representational accuracy, coherence with background knowledge, and assessment of the domain of validity. In some cases, however, we might know enough about the domain that we can give a sufficiently confident assessment of machine learning models' coherence with background knowledge. In those cases, they can explain the phenomena they are trained to predict. Yet, what kind of explanations can be obtained if the processes the models implement remain unclear?

Meskhidze (2023) tries to provide more substantive answers here. She argues that machine learning models in cosmology (predicting cosmological parameters in large simulations of the formation of galaxies) answer some why-questions, but do not help us understand "why phenomena of this general type occur across a variety of circumstances" (Meskhidze 2023, 1901). The reason is their lack of physical representations; they do not adhere to physical laws and so are not suited to explain such questions about the unfolding of physical processes. This is not a problem, though, for such machine learning models to help us understand "why, for example, our universe has the particular distribution of matter it does. By filling out the parameter space of interest, such methods can point cosmologists to the relevant values of the cosmological parameters that led to a particular distribution of matter" (Meskhidze 2021, 1906). The argument seems to be that if the outputs of the machine learning models correspond to the actual values, then this can be explanatory of the actual distribution of matter. However, scientific explanations are typically thought to require a covering rule or mechanism sketch. The machine

learning model does not seem to provide that overarching process, which is instead given by physics-based N-body models. As such, the machine learning model does not seem to answer the question of why our universe has the particular distribution of matter that it does on its own. It thus remains unclear what the explanatory value of the machine learning models is exactly.

Despite widespread optimism, no clear answers have emerged on how machine learning models lead to (novel) explanations, even if the link uncertainty is reduced. At the same time, the pessimists might be too hasty to dismiss the extraction of causal relations from machine learning models, as there is a burgeoning literature connecting causal inference to machine learning (Pearl 2019). Buijsman (2023) connects this literature to machine learning techniques for causal inference to argue that in a few specific cases we can get (causal) scientific explanations from machine learning models. However, he also argues that this is unlikely to work for predictive machine learning models due to inherent biases in these models. Furthermore, causal accounts of explanation are not the only option. Other epistemic accounts of explanation are likewise viable; for example, Durán (2017, 2021) approaches scientific explanation and machine learning from a unificationist perspective. Such alternative accounts deserve more attention in the debate on scientific explanations from machine learning. The central challenge of formulating how explanations arise from machine learning models (if at all), remains an open question and calls for both a broader look at explanations and more in-depth case studies.

Let us finally note that the current debate on understanding machine learning largely happens in light of explanation. This is either because explanations are seen by many as a one-solution-fits-all (e.g., it reduces opacity, increases transparency, provides trustworthy machine learning, and adds to our understanding of the system) or because it is the standard philosophical pathway to understanding. Take, for instance, the objections raised by Räz and Beisbart (2022) to Sullivan's uncertainty link. To these authors, Sullivan's view depends on which notion of understanding is at play, and a strong notion would require explanatory understanding. Although they do not adopt a specific definition of explanatory understanding, they accept de Regt's (2017) and Khalifa's (2017) as suitable interpretations for their purposes. In this context, the overall strategy of Räz and Beisbart consists of showing that understanding ML comes in close connection with explanations. But not just any form of explanation. In particular, Sullivan's how-explanations strike them as unconvincing: "She writes that the deep patient model can answer the question of 'how it is possible to predict disease development for a range of diseases'" (Sullivan 2022, 123). As pointed out by Räz and Beisbart, "This is not a request for a how-possibility explanation of phenomena in the target system, it is a question about the possibility of predictive modeling itself" (2022).

Some authors have taken a somewhat different path in the connection between explanation and understanding. Páez (2019), for instance, claims that the search for explainable AI must be formulated in terms of the broader project of offering a pragmatic and naturalistic account of understanding. The result is the same: the analysis of explanations is in light of understanding. But is there a way to address understanding without resorting to explanation (and vice versa)? Räz and Beisbart think so. They suggest that machine learning can produce some degree of objectual understanding, here taken to be:

> the understanding of a domain of things; it is often taken to imply some knowledge of this domain and the grasp of connections between items in the domain. These connections may be explanatory, but need not be; they may be merely logical or probabilistic.
>
> *(Räz and Beisbart 2022)*

Examples of objectual understanding have been discussed in the philosophical literature. Gijsbers (2013) shows that some classifications, such as those used in biology, can effectively enhance our understanding of, say, species without providing explanations. Based on these, Räz and Beisbart suggest that "ML models can lead to some objectual understanding, e.g., by establishing correlations, or by simply adding to knowledge of a domain of things" (Räz and Beisbart 2022).

4. Conclusion

What are the epistemic implications of machine learning models in the sciences? In the case of neuroscience, these epistemic implications are fairly clear. Neural networks, a type of machine learning model, can be seen as representing (parts of) the brain, and elements of neural networks can be linked to elements of biological neurons and synapses. Questions remain on the limitations of neural networks as models of the brain, e.g., due to their limited ability to generalize, but it is clear that they play a role in understanding the functioning of the brain.

When Machine Learning functions as a tool, in other sciences, its contribution to understanding is far less clear. Since machine learning models do not contain physical representations, they are harder to link to the actual situation they model. Furthermore, their epistemic opacity makes it difficult to extract causal relations and even to determine the reliability and robustness of their predictions. As a result, it is unclear when scientists are justified to believe the predictions made by such models, and more work is needed on specifying exactly what conditions hold for justification in these contexts. Furthermore, it is unclear whether and what explanations (and understanding) can be gained from machine learning models. The discussion so far has focused on causal accounts of explanation but has not yet yielded examples of causal explanations that are clearly obtained from the machine learning model. Both a broader look at accounts of explanations and more detailed case studies are needed to determine the explanatory role of machine learning models in science. Such models are here to stay due to their benefits of higher empirical accuracy and lower computational costs, as the range of case studies has shown.

Acknowledgments

Juan M. Durán has received support from the EU program under the scheme "INFRAIA 2020-2024-SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics," grant agreement 871042. He has also received support from the EU program under the scheme "ICT48 Humane AI Net," grant agreement 952026. Their support is gratefully acknowledged.

References

Agarwal Shankar, Filipe B. Abdalla, Hume A. Feldman, Ofer Lahav, and Shaun A. Thomas. 2012. "PkANN - I. Non-Linear Matter Power Spectrum Interpolation through Artificial Neural Networks." *MNRAS* 424: 1409–1418.

Alcorn, Michael A., Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. 2019. "Strike (with) a Pose: Neural Networks are Easily Fooled by Strange Poses of Familiar Objects." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 4845–4854.

Beisbart, Claus. 2017. "Advancing Knowledge through Computer Simulations? A Socratic Exercise." In *The Science and Art of Simulation I*, edited by Michel Resch, Andreas Kaminski, and Petra Gehring, 153–174. Berlin: Springer.

———. 2021. "Opacity Thought Through: on the Intransparency of Computer Simulations." *Synthese* 199: 11643—11666.

Buckner, Claus. 2018. "Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks." *Synthese* 195(12): 5339–5372.

Buijsman, Stefan. (2023). Causal scientific explanations from machine learning. *Synthese*, 202(6), 202.

Das, Arun, and Paul Rad. 2020. "Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey." *arXiv preprint* arXiv:2006.11371.

de Regt, Henk W. 2017. *Understanding Scientific Understanding*. New York: Oxford University Press.

Durán, Juan M. 2017. "Varying the Explanatory Span: Scientific Explanation for Computer Simulations." *International Studies in the Philosophy of Science* 31(1): 27–45.

———. 2021. "Dissecting Scientific Explanation in AI (sXAI): A Case for Medicine and Healthcare." *Artificial Intelligence* 297: 103498.

Durán, Juan M., and Nico Formanek. 2018. "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28(4): 645–666.

Frigg, Roman, and Julian Reiss. 2009. "The Philosophy of Simulation: Hot New Issues or Same Old Stew?" *Synthese* 169(3): 593–613.

Gijsbers, Victor. 2013. "Understanding, Explanation, and Unification." *Studies in History and Philosophy of Science Part A* 44.3: 516–522.

Güçlü, Umut, and Marcel J. van Gerven, M. A. 2017. "Modeling the Dynamics of Human Brain Activity with Recurrent Neural Networks." *Frontiers in Computational Neuroscience* 11: 7.

Humphreys, Paul W. 2009. "The Philosophical Novelty of Computer Simulation Methods." *Synthese* 169(3): 615–626.

Jebeile, Julie, Vincent Lam, and Tim Räz. 2021. "Understanding Climate Change with Statistical Downscaling and Machine Learning." *Synthese* 199(1): 1877–1897.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature*, 596: 583–589.

Kawamleh, Suzanne. 2021. "Can Machines Learn How Clouds Work? The Epistemic Implications of Machine Learning Methods in Climate Science." *Philosophy of Science* 88(5): 1008–1020.

Khalifa, Kareem. 2017. *Understanding, Explanation, and Scientific Knowledge*. Cambridge: Cambridge University Press.

Knüsel, Benedikt, and Christoph Baumberger. 2020. "Understanding Climate Phenomena with Data-Driven Models." *Studies in History and Philosophy of Science Part A* 84: 46–56.

López-Rubio, Ezequiel. 2018. "Computational Functionalism for the Deep Learning Era." *Minds and Machines* 28(4): 667–688.

López-Rubio, Ezequiel, and Emanuelle Ratti. 2021. "Data Science and Molecular Biology: Prediction and Mechanistic Explanation." *Synthese* 198(4): 3131–3156.

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

Miłkowski, Marcin. 2013. *Explaining the Computational Mind*. Cambridge: MIT Press

Páez, Andrés. 2019. "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)." *Minds and Machines* 29: 441—59.

Pearl, Judea. 2019. "The Seven Tools of Causal Inference with Reflections on Machine Learning." *Communications of the ACM* 62(3): 54–60.

Perrakis, Anastassis, and Titia K. Sixma. 2021. "AI Revolutions in Biology: The joys and Perils of AlphaFold." *EMBO Reports* 22(11): e54046.

Piccinini, Gualtiero. 2010. "The Mind as Neural Software? Understanding Functionalism, Computationalism, and Computational Functionalism." *Philosophy and Phenomenological Research* 81(2): 269–311. https://doi.org/10.1111/j.1933-1592.2010.00356.x

Piccinini, Gualtiero, and Craver, Carl. 2011. Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*, 283-311.

Rasp, Stephan, Michael S. Pritchard, and Pierre Gentine. 2018. "Deep Learning to Represent Subgrid Processes in Climate Models." *PNAS* 115(39): 9684–9689.

Räz, Tim, and Claus Beisbart. 2022. "The Importance of Understanding Deep Learning." *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y

Rosenfeld, Amir, Richard Zemel, and John K. Tsotsos. 2018. "The Elephant in the Room." *arXiv preprint*. arXiv:1808.03305.

Schmidt, Jonathan, Mario R. Marques, Silvana Botti, and Miguel A. L. Marques. 2019. "Recent Advances and Applications of Machine Learning in Solid-State Materials Science." *npj Computational Materials* 5(1): 1–36.

Skolnick, Jeffrey, Mu Gao, Hongyi Zhou, and Suresh Singh. 2021. "AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function." *Journal of Chemical Information and Modeling* 61(10): 4827–4831.

Srećković, Sanja, Andrea Berber, and Nenad Filipović. 2022. "The Automated Laplacean Demon: How ML Challenges Our Views on Prediction and Explanation." *Minds and Machines* 32: 159–183.

Stinson, Catherine. 2018. "Explanation and Connectionist Models." In *The Routledge Handbook of the Computational Mind*, edited by Matteo Colombo and Mark Sprevak, 120–133. New York: Routledge.

Sullivan, Emily. 2022. "Understanding from Machine Learning Models." *British Journal for the Philosophy of Science* 73(1): 109–133.

Symons, John, and Ramón Alvarado. 2019. "Epistemic Entitlements and the Practice of Computer Simulation." *Minds and Machines* 29(1): 37–60.

Teichroew, Daniel, and John F. Lubin. 1966: "Computer Simulation Discussion of the Technique and Comparison of Languages." *Communications of the ACM* 9(10): 723–741.

Zhuang, Chengxu, Siming Yan, Aran Nayebi, Martin Schrimpf, Michel C. Frank, James J. DiCarlo, and Daniel Yamins. 2021. "Unsupervised Neural Network Models of the Ventral Visual Stream." *PNAS* 118(3): e2014196118.