DOI: 10.1017/psa.2023.8

This is a manuscript accepted for publication in *Philosophy of Science*.

This version may be subject to change during the production process.

Deep Learning Opacity in Scientific Discovery

Eamon Duede

Department of Philosophy

Committee on Conceptual & Historical Studies of Science

Pritzker School of Molecular Engineering

Knowledge Lab

University of Chicago

eduede@uchicago.edu

December 12, 2022

Abstract

While philosophers have focused on the epistemological and social challenges of using Artificial Intelligence in science, scientists have focused on the opportunities. I argue that this disconnect between philosophical pessimism and scientific optimism is driven by failures to critically examine the practice of AI-infused science. To appreciate the epistemic justification for AI-powered breakthroughs, philosophers must analyze the role of deep learning as part of a wider process of discovery. I demonstrate the importance of this with two cases from the scientific literature, and show that epistemic opacity need not diminish AI's capacity to lead scientists to significant and justifiable breakthroughs.

1 Introductory

The recent boom in optimism for the use of deep learning (DL) and artificial intelligence (AI) in science is due to the astonishing capacity of deep neural networks to facilitate discovery [DeVries et al., 2018], overcome the complexity of otherwise intractable scientific problems [Senior et al., 2020], as well as to both emulate and outperform experts on routine [Chen et al., 2014], complex, or even humanly impossible [Degrave et al., 2022] tasks. In fact, nearly every empirical discipline has already undergone some form of transformation as a result of developments in and implementation of deep learning and artificial intelligence [Stevens et al., 2020]. To scientists and science funding agencies alike, artificial intelligence both promises and has already begun to revolutionize not only our science, but our society, and quality of life.

Yet, someone reading the recent philosophical literature on deep learning might be forgiven for concluding that doing good science with deep neural networks must be exceptionally challenging, if not impossible. This is because philosophers have, of late, focused not on the enormous potential of DL and AI, but on a number of important epistemological challenges that arise from the uninterpretability of deep neural networks (DNNs). Given that DNNs are epistemically opaque [Creel, 2020, Humphreys, 2009, Zerilli, 2022, Lipton, 2018], it is, in many instances, impossible to know the high-level, logical rules that govern how the network relates inputs to outputs. It is argued that this lack of transparency severely limits scientists' ability to form explanations for and understanding of why neural networks make the suggestions that they do [Creel, 2020, Zerilli, 2022, Sullivan, 2019]. For instance, Creel states that "access to only observable inputs and outputs of a completely opaque black-box system is not a sufficient basis for explanation[.]" [Creel, 2020, pg.573]. So, this presents an obvious epistemic challenge when explanations of neural network logic are required to justify claims or decisions made on the basis of their outputs.

As a result, reading the recent philosophical literature can leave one wondering on what basis (beyond mere inductive considerations) neural network outputs can be justified [Boge, 2021]. While lack of interpretability is of particular concern in high-stakes decision-making settings where accountability and value-alignment are salient (e.g., medical diagnosis and criminal justice) [Falco et al., 2021, Hoffman, 2017], the opacity of deep learning models may also be of concern in basic research settings where explanations and understanding represent central epistemic virtues and often serve as justificatory credentials [Khalifa, 2017]. Even if inductive considerations such as the past success of the model on out of sample data can help to raise confidence in the scientific merit of its outputs, in general, without additional justification, it is unclear how scientists can ensure that results are consistent with the epistemic norms of a given discipline.[1] Or, so we are led to conclude.

There is, then, a sharp contrast between the relative optimism of scientists and policymakers on the one hand, and the pessimism of philosophers on the other, concerning the use of deep learning methods in science. This disconnect is due, I believe, to a failure on the part of philosophers to attend to the full range of ways that deep learning is actually used in science.[2] In particular, while philosophers are right to examine and raise concern over epistemological issues that arise as a result of neural network opacity, it is equally

[1] To that end, scientists and philosophers have turned to various nascent approaches under the heading of Explainable AI (XAI). Nevertheless, the presumption is still that, absent explanation, justification for network outputs will be hard to come by.

[2] Though, I also believe that scientists routinely underestimate the epistemological challenge presented by DNN opacity.

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

important to step back and analyze whether these issues do, in fact, arise in practice and, if so, in what contexts and under what conditions.

In this paper, I argue that epistemological concerns due to neural network opacity will arise chiefly when network outputs are treated as scientific claims that stand in need of justification (e.g., treated as candidates for scientific knowledge, or treated as the basis for high-stakes decisions). It is reasonable to think that this must happen quite a bit, particularly outside of scientific settings. After all, the promise of deep learning is the rapid discovery of new knowledge. Of course, philosophers are correct that, if neural network outputs are evaluated in, what has often been referred to as, the "context of justification", then access to the high-level logic of the network (e.g., interpretability) will, in most cases, be required for validation. While this certainly happens, I will show that scientists can make breakthrough discoveries and generate new knowledge utilizing fully opaque deep learning without raising any epistemological alarms. In fact, scientists are often well aware of the epistemological limitations and pitfalls that attend the use of black-box methods. But, rather than throw up their arms and embrace a form of pure instrumentalism (or worse, bad science), they can carefully position and constrain their use of deep learning outputs to what philosophers of science have called the "context of discovery".

The paper proceeds as follows: In *Section 2*, drawing on recent philosophy of science, I explain the relevant sense in which deep learning models are opaque. In *Section 3*, I sketch the epistemological distinction between treating opaque model outputs as standing in need of justification and of treating such out puts as *part* of a wider process of discovery. In *Section 4*, I present two cases which demonstrate the way in which researchers can make meaningful scientific discoveries by means of deep learning. Yet, in both cases, the findings do not rely on interpretability or accuracy for their justification.

2 Deep Learning Opacity

Deep learning is a machine learning technique based on artificial neural networks that is widely used for prediction and classification tasks. The goal of deep learning is to automate the search for a function $\hat{f}$ that approximates the true function f that generates observed data. The fundamental assumption that motivates the use of deep learning is that f is in the set of functions $\mathcal{F}$ *representable* by a neural network given some particular architecture and (hyper)parameterization. Of course, for any given parameterization k, we have no way of knowing *a priori* whether $f \in \mathcal{F}_k$. However, deep neural networks are universal approximators [Hornik et al., 1989], so the assumption is at least principled.

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

Like most regression tasks, the trained model $\hat{f}$ is arrived at by iteratively minimizing a loss function $\mathcal{L}$ through back-propagation of error gradients and updating all weights in the network such that the risk R over the training distribution P is minimized for $R_p(\hat{f}) := \mathbb{E}_{(X,Y)\sim P}[\mathcal{L}(\hat{f}(X), Y)]$ where X and Y are sets of inputs and outputs. It is assumed that $\hat{f}$ is low risk over the distribution used for training if it performs well on a randomly selected iid test set from P. As a result, a highly accurate model is expected to perform well on out of training sample data which, in turn, provides a high degree of inductive support for confidence in the accuracy of its outputs. Nevertheless, while inductive considerations are common for assessing the merit of claims in science, they typically fall short of the justificatory standard of most disciplines.

From the above, it should be clear that there is a straightforwardly mathematical sense in which deep neural networks are *fully transparent*. All weights on all connections across the network, billions as there may be, are both available to inspection and computationally tractable. However, while formally precise, neural network logic is largely semantically unintelligible. That is, the mathematical expression of a fully trained neural network model cannot, in general, be given an intelligible interpretation in terms of the target system such that one can understand or comprehend how the parts interact and contribute to the networks' outputs.

[Zerilli, 2022] describes the opacity of deep learning models (DLMs) by bringing out the distinction between "Tractability", "Intelligibility", and "Fathomability", a distinction echoed in [Lipton, 2018]. Here, the idea is that any working machine learning model is tractable in so far as it can be run on a computer. However, intelligibility comes in degrees that are modulated by model fathomability. Fathomability is understood to be the extent to which a person can understand, straight away, how the model relates features to produce outputs. As a result, the more complex a model (e.g., increased dimensionality, extreme nonlinearities, etc.), the less fathomable it becomes. Many highly complex but linear models (e.g., random forests) remain "intelligible" in so far as all of the relationships between elements of the model can, in principle, be semantically deciphered even though the model as a whole (its overall decision logic) remains unfathomable due to complexity.

Zerilli's three aspects of epistemic access to neural network logic mirror [Creel, 2020]'s three levels or granular scales of transparency. For Creel, the transparency of a complex, computational model can be assessed "Algorithmically", "Structurally", and at "Runtime". Most relevant to the issues of this paper are algorithmic and structural transparency. For Creel, a model is algorithmically transparent if it is possible to establish which high-level, logical rules (e.g., which *algorithm*) govern the transformation

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

of input to output. In the case of a deep neural network, it is not possible to know which algorithm is implemented by the network precisely because the algorithm is developed autonomously during training. As a result, DNNs also lack what Creel calls "structural" transparency in that it is not clear how the distribution of weights and (hyper)parmeterization of the neural network implements (realizes) the algorithm that it has learned. Therefore, for Creel, DNNs are opaque —neither "fathomable" nor "intelligible" in Zerilli's sense.[3] Following [Humphreys, 2004, Humphreys, 2009], a process is said to be *epistemically* opaque when it is impossible for a scientist to know all of the factors that are epistemically relevant to licensing claims on the basis of that process, where factors of 'epistemic relevance' include those falling under Creel's algorithmic and structural levels and Zerilli's intelligibility and fathomability criteria. As such, DLMs are "epistemically opaque".

3 Discovery and Justification with Deep Learning

When it comes to justifying belief or trust in the outputs of deep learning models, their epistemic opacity is straightforwardly problematic. This is due to the fact that it is not possible to evaluate all of the epistemically relevant factors that led to the output. In high-stakes settings such as medical diagnosis, where the output of an epistemically opaque model forms the basis of a decision, a decision maker's inability to explain why the model prompts the decision that it does (and not, say, some other decision) can raise reasonable doubt as to whether the decision is, in fact, justified. Here is Creel on why we should strive for transparency:

> "I claim that we should [strive for transparency] because scientists, modelers, and the public all require transparency and because it facilitates scientific explanation and artifact detection [...] Descriptively, the scientists who use the [epistemically opaque] systems to investigate, the modelers and computer scientists who create the systems, and the nonscientist citizens who interact with or are affected by the systems all need transparency." [Creel, 2020, pg.570]

Why might scientists (and others) all *need* and *require* transparency? The reasons Creel and most philosophers[4] concerned with the epistemology of deep learning give are that, without transparency, scientists are unable to understand the outputs of their models, are powerless to explain why the models perform the way they do, cannot provide justification for the decisions they make on the basis of the model output, are uncertain whether and to what extent the models reflect our values —on and on. What all of these

[3] Both Creel and Zerilli draw on the computational concepts of understanding in information processing systems developed by David Marr [Marr, 1977].

[4] There are exceptions including [Lenhard and Winsberg, 2010, Hooker and Hooker, 2018].

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

reasons have in common is a commitment to the idea that neural network transparency is *epistemically essential* to effectively use and gain knowledge from powerful artificial intelligence applications in scientific and societal settings [Gunning et al., 2019].

However, I argue that there are many cases in which the epistemic opacity of deep learning models is *epistemically irrelevant* to justifying claims arrived at with their aid. That is, it is justifiably possible to effectively use and gain scientific knowledge from epistemically opaque systems without sacrificing any justificatory rigor at all. In fact, scientists routinely achieve breakthroughs using deep learning that far exceed what they would have been able to do without such systems, while neither needing nor requiring transparency to justify their findings. This, I argue, can be readily observed when considering how epistemically opaque models can figure in the generation of findings.

Philosophers of science have, at least since [Reichenbach, 1938], (and, later, [Popper, 2002]) drawn a logical distinction between what has typically been referred to as the contexts of "justification" and "discovery". I argue that concerns about opacity are only epistemically relevant in settings where the outputs of epistemically opaque models are treated as candidates for scientific knowledge in their own right, that is, treated as claims that stand in need of justification. In these settings, neural network outputs are treated as the end result of an investigation (e.g., as findings in their own right) and, as such, fall within the "context of justification". However, in the context of justification, rigorous evaluation of the reasons that support findings is required to justify them. If those reasons are the internal logic of a neural network, then this kind of evaluation is blocked by neural network opacity.

Be that as it may, the outputs of epistemically opaque models need not be treated in this way. Rather, they can serve as aspects or parts of a *process of discovery*. While the process ultimately leads to claims that stand in need of justification, the part played by an opaque model in that process can, itself, be epistemically insulated from the strong sort of evaluation that is applied to findings in the context of justification. In this way, neural network outputs can serve to facilitate discovery without their outputs or internal logic standing in need of justification. That is, neural network outputs that serve as parts of a process of discovery (similar to abduction [Duede and Evans, 2021, Hanson, 1965] and problem-solving heuristics [Wimsatt, 2007, Simon, 1973]) can be treated as situated in the "context of discovery".[5]
In the context of discovery, the outputs of neural networks can be used to guide attention and scientific intuition toward more promising hypotheses but do not, themselves, stand in need of justification. Here, outputs of opaque models serve to provide reasons to or

[5]For a more nuanced view of the context of discovery see, for example, [Laudan, 1981].

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

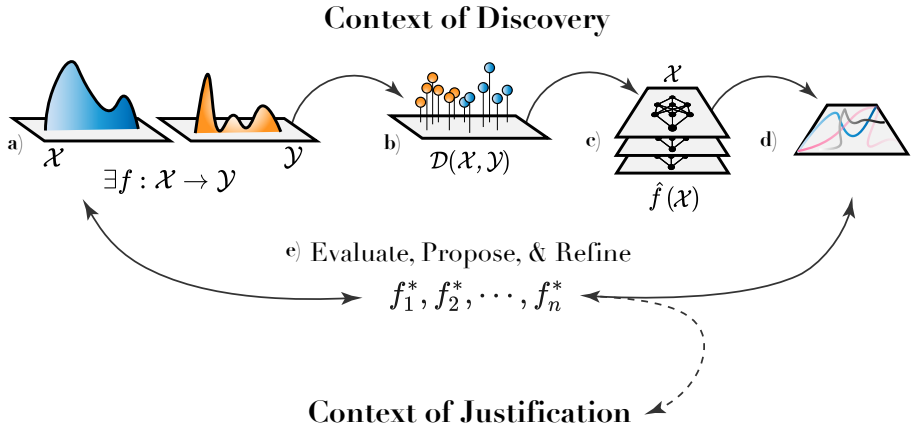

Figure 1: The confinement of epistemically opaque, neural network outputs to the context of discovery. **a)** posit the existence of some theory $\exists f$ that connects two phenomena; **b)** generate a dataset D that represents the assumed connection; **c)** train a deep learning model to learn a function that approximates the posited theory; **d)** examine the behavior of $\hat{f}$; **e)** iteratively evaluate **(b)-(d)**, formulate, and refine hypotheses f_i^* connecting phenomena. Justify f^* by means distinct from those used to produce it.

evidence for pursuit of particular paths of inquiry over others (see: Figure 1). As such, they both provide and are subject to forms of preliminary appraisal, but, as the cases in *Section 4* will bring out, the mere inductive support DLMs provide is epistemically sufficient to guide pursuit.

4 Justified Discoveries Using Deep Learning

In this section, I present two cases that demonstrate the claim that the epistemic opacity of DLMs can be epistemologically irrelevant for justifying the very scientific findings facilitated by their use.

4.1 Case 1: Guiding Mathematical Intuition

Here, I consider a case from low-dimensional topology in which researchers use deep learning to guide mathematical intuition concerning the relationship between two classes of properties of knots. Knots are interesting topological objects because the relationships between their properties are poorly understood, while their connections to other fields within mathematics are plausible but unproven.

In [Davies et al., 2021], mathematicians seek to discover and prove a conjecture that establishes a mathematical connection between known geometric and algebraic properties of knots in $\mathbb{R}^3$. In particular, the aim is to establish that hyperbolic invariants of knots (e.g., geometric properties of knots that are identical for all equivalent knots) and alge-

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

braic invariants of knots are connected. While the possibility of this connection had been imagined, its plausibility had not been established empirically and certainly not proved mathematically. As a result, mathematical intuition concerning a possible connection was too vague for genuine insight to emerge.

The deep learning approach in [Davies et al., 2021] begins by *imagining* that the geometric invariants X of a given knot K are, in fact, connected to that knot's algebraic invariants Y. The algebraic invariant $\sigma \in Y$ called the 'signature' is known to represent information about a given knot's topology. The researchers hypothesized that a knot's signature $\sigma(K)$ is provably related to its hyperbolic invariants $X(K)$ such that there exists some function f such that $f(X(K)) = \sigma(K)$ (Figure 1a). They then constructed a dataset of observed hyperbolic invariants and signatures for individual knots and trained a DLM to predict the latter from the former (Figure 1b). If the resulting DLM (Figure 1c) achieves an accuracy better than chance on a holdout set, then this provides researchers with reasons to expect that *some* mathematical relationship must obtain between $X(K)$ and $\sigma(K)$. Importantly, however, accepting this claim ultimately serves no role in proving the conjectured, mathematical relationship.

In fact, the initial DLM achieved an accuracy of roughly 0.78, giving researchers high confidence that a connection between hyperbolic invariants and algebraic invariants obtains. While the established plausibility of the connection might be sufficient to find a promising conjecture, it is possible to isolate the various geometric invariants most responsible for the accuracy of $\hat{f}$. Specifically, by quantifying the change in the gradient of the loss function with respect to each of the individual geometric invariants,[6] it is possible to guide attention to a subset of hyperbolic invariants to consider when formulating a conjecture (Figure 1d). In particular, three such invariants are found to be most responsible for DLM accuracy in predicting signatures: the real and imaginary parts of the meridional translation μ and the longitudinal translation λ.

From these elements, mathematicians used their intuition to formulate an initial conjecture (Figure 1e) that relates $\mu(K)$ and $\lambda(K)$ to $\sigma(K)$ by means of a novel, conjectured property which they call the natural slope ($Re(\lambda/\mu)$) where Re denotes the real part of the meridional translation) of K. Using common computational techniques, corner cases were constructed that violate the initial conjecture, which was, in turn, refined into the following theorem.

[6]The saliency quantity r is calculated by averaging the gradient of the cross-entropy loss function with respect to each geometric invariant x_i over all training examples such that $\mathrm{r_i} = \frac{1}{|X|} \sum \mathbf{x} \in X \left| \frac{\partial L}{\partial \mathbf{x}_i} \right|$

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

Theorem: There exists a constant c such that, for any hyperbolic knot K,

$$|2\sigma(K) - \text{slope}(K)| \leq c\,\text{vol}(K)\,\text{inj}(K)^{-3} \tag{1}$$

Let's call the above theorem T. That T is true is provable. As a result, justification for belief in the truth of T does not lie in any empirical considerations. Nor does it rely on any facts about how the conjecture was arrived it. That is, the proof for T does not depend on any of the steps that were taken for its discovery. Its justificatory status is not diminished by the fact that a number of assumptions were made in the process of its discovery nor that an opaque model aided in the decision to take seriously the connection between hyperbolic and algebraic properties.

One might object to the claim that the opacity of the network in this case was epistemically irrelevant. After all, the gradient based saliency method used to isolate the contribution to accuracy of the various inputs might be viewed as an interpretive step. Yet, it is important to note that this saliency procedure was applied to the input layer which is, necessarily, transparent to begin with. Ultimately, then, what this saliency method adds is computational expediency as, alternatively, it was possible to use a simple combinatorial approach to iteratively search through the input space for the most predictive properties. In this way, saliency adds nothing of epistemic relevance to the process that would not have been possible without it. Moreover, saliency is not required for justification. Nevertheless, as we will see, *Case 2* is an example in which no such interpretive step is taken.

4.2 Case 2: Deep Learning for Theory Improvement

This case demonstrates the use of deep learning to dramatically improve our understanding of the geophysics of earthquakes by providing researchers good reasons to consider integrating known geophysical properties into existing theory. Consistent with the central claim of this paper, the neural network itself, while opaque, neither contributes nor withholds anything of epistemic importance to the justificatory credentials of the reworked theory.

The geophysics of earthquakes is poorly understood. In [DeVries et al., 2018], scientists seek to improve theory that describes the dynamics relating aftershocks to mainshocks. The best available theoretical models of aftershock triggering dynamics correctly predict the location of an aftershock with an $AUC = 0.583$. As the authors point out, while "the maximum magnitude of aftershocks and their temporal decay are well described by empirical laws [...] explaining and forecasting the spatial distribution of aftershocks is

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

more difficult." [DeVries et al., 2018, pg.632]

To overcome this difficulty, scientists turn to deep learning to evaluate whether and to what extent it is possible to functionally relate mainshock and aftershock locations. After all, if an essentially stochastic process relates locations, then perhaps the extant theory is empirically adequate. As in *Case 1*, researchers begin by imagining that aftershock locations are a function f of mainshocks and seek to find an approximation of that function $\hat{f}$ (Figure 1a). To operationalize this, they construct a dataset of mainshock and aftershock events by representing the planet as a collection of 5km^3 tiles (Figure 1b). A tile is experiencing a mainshock, an aftershock, or no shock at any given moment. For the purposes of relating events, it is sufficient to treat the prediction as a simple, binary classification task (Figure 1c). Given an input (mainshock parameter values and affected tiles), the task is to classify every terrestrial tile as either 'aftershock' or 'not aftershock'.

The fully trained DLM correctly forecasts the locations of aftershocks between one second to one year following a mainshock event with an $AUC = 0.85$, significantly outperforming theory. This gives researchers good reason to take seriously the possibility of further improving theory (Figure 1d). Yet, this reason is not implicated in nor relevant to justifying the reworked theory.

The DLM outputs a probability distribution of 'aftershock' over tiles. The researchers compare this distribution to the one predicted by extant theory. Surprisingly, they observe that the probability of aftershock within a certain radius of a mainshock assigned by theory is largely uncorrelated with the probabilities assigned by the DLM. At this point, the DLM's predictive power is treated as evidence that theory can be significantly improved thereby guiding the process of discovery.

Given the network's high accuracy, it is reasonable to assume that an improved theory would more closely resemble the observed probability distribution generated by the network. By iteratively sweeping through known geophysical properties and correlating them with DLM distributions, they find that three parameters (maximum change in shear stress, the von Mises yield criterion, and aspects of the stress-change tensor), that had not been considered by geophysicists as relevant, in fact explain nearly all of the variance in predictions generated by the neural network, thereby providing novel physical insight into the geophysics of earthquakes (Figure 1e).

In this case, a fully opaque DLM has profound implications for our theoretical understanding of earthquake dynamics. Namely, the ability to accurately predict phenomena orients scientific attention to empirical desiderata necessary for more accurate theory

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

building. Moreover, it is epistemically irrelevant to justifying the improved theory that we cannot verify whether and how any of the geophysical quantities that were determined to be of relevance are, in fact, represented in the network. This is because it is not the network's predictions that stand in need of justification but, rather, the theory's itself. The reworked theory is justified in ways that are consistent with the norms of the discipline —it relates known geophysical properties in ways that are consistent with first principles, it aids in the explanation and understanding of aftershock dynamics, and it outperforms extant theory in prediction. Yet, none of this depends on the neural network that was used to lead attention to relevant revisions of the theory for justification.[7]

5 Discussion

What I hope to have shown in this paper is that, despite their epistemic opacity, deep learning models can be used quite effectively in science, not just for pragmatic ends but for genuine discovery and deeper theoretical understanding, as well. This can be accomplished when DLMs are used as *guides* for exploring promising avenues of pursuit in the context of discovery. In science, we want to make the best conjectures and pose the best hypotheses that we can. The history of science is replete with efforts to develop processes for arriving at promising ideas. For instance, thought experiments are cognitive devices for hypothesis generation, exploration, and theory selection. In general, we want our processes of discovery to be as reliable or trustworthy as possible. But, here, inductive considerations are, perhaps, sufficient to establish reliability. After all, the processes by which we arrive at our conjectures and hypotheses do not typically serve also to justify them. While philosophers are right to raise epistemological concerns about neural network opacity, these problems primarily concern the treatment and use of deep learning outputs as findings in their own right that stand, as such, in need of justification which (as of now) only network transparency can provide. Yet, when DLMs serve the more modest (though no less impactful) role of guiding science in the context of discovery, their capacity to lead scientists to significant breakthroughs is in no way diminished.

6 Acknowledgements

Revised versions of this manuscript benefited greatly from conversations with Kevin Davey, James Evans, Ian Foster, Tyler Millhouse, Tom Pashby, Bill Wimsatt, and participants of the University of Texas at Austin's Philosophy of Biology Circle Working Group and the *Philosophy of Science Association's* 28^{th} Biennial Meeting. This work was

[7]Note that no relevant geophysical properties are in the input or output space of the model. So, none of these properties are explicitly modeled. As a result, we have what [Wimsatt, 2007] might call a false model that, nevertheless, leads to "truer theory".

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press

supported by the US *National Science Foundation* #2022023 NRT-HDR: AI-enabled Molecular Engineering of Materials and Systems (AIMEMS) for Sustainability.

References

[Boge, 2021] Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, pages 1–33.

[Chen et al., 2014] Chen, Y., Lin, Z., et al. (2014). Deep learning-based classification of hyperspectral data. *IEEE Journal of Selected topics in applied earth observations and remote sensing*, 7(6):2094–2107.

[Creel, 2020] Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4):568–589.

[Davies et al., 2021] Davies, A., Veličković, P., et al. (2021). Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887):70–74.

[Degrave et al., 2022] Degrave, J., Felici, F., et al. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419.

[DeVries et al., 2018] DeVries, P. M., Viégas, F., et al. (2018). Deep learning of aftershock patterns following large earthquakes. *Nature*, 560(7720):632.

[Duede and Evans, 2021] Duede, E. and Evans, J. (2021). The social abduction of science. *arXiv preprint arXiv:2111.13251*.

[Falco et al., 2021] Falco, G., Shneiderman, B., et al. (2021). Governing ai safety through independent audits. *Nature Machine Intelligence*, 3(7):566–571.

[Gunning et al., 2019] Gunning, D., Stefik, M., et al. (2019). Xai—explainable artificial intelligence. *Science Robotics*, 4(37):eaay7120.

[Hanson, 1965] Hanson, N. R. (1965). *Patterns of discovery: An inquiry into the conceptual foundations of science*. CUP Archive.

[Hoffman, 2017] Hoffman, R. R. (2017). A taxonomy of emergent trusting in the human–machine relationship. *Cognitive Systems Engineering*, pages 137–164.

[Hooker and Hooker, 2018] Hooker, G. and Hooker, C. (2018). Machine learning and the future of realism. *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 9(1):174–182.

[Hornik et al., 1989] Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

[Humphreys, 2004] Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.

[Humphreys, 2009] Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3):615–626.

[Khalifa, 2017] Khalifa, K. (2017). *Understanding, explanation, and scientific knowledge*. Cambridge University Press.

[Laudan, 1981] Laudan, L. (1981). Why was the logic of discovery abandoned? In *Science and hypothesis*, pages 181–191. Springer.

[Lenhard and Winsberg, 2010] Lenhard, J. and Winsberg, E. (2010). Holism, entrenchment, and the future of climate model pluralism. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 41(3):253–262.

[Lipton, 2018] Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57.

[Marr, 1977] Marr, D. (1977). Artificial intelligence—a personal view. *Artificial Intelligence*, 9(1):37–48.

[Popper, 2002] Popper, K. (2002). *Popper: The logic of scientific discovery*. Routledge Classics New York, NY.

[Reichenbach, 1938] Reichenbach, H. (1938). Experience and prediction: An analysis of the foundations and the structure of knowledge.

[Senior et al., 2020] Senior, A. W., Evans, R., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

[Simon, 1973] Simon, H. A. (1973). Does scientific discovery have a logic? *Philosophy of science*, 40(4):471–480.

[Stevens et al., 2020] Stevens, R., Taylor, V., et al. (2020). Ai for science. Technical report, Argonne National Lab.(ANL), Argonne, IL (United States).

[Sullivan, 2019] Sullivan, E. (2019). Understanding from machine learning models. *British Journal for the Philosophy of Science*.

[Wimsatt, 2007] Wimsatt, W. C. (2007). *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Harvard University Press.

[Zerilli, 2022] Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science*, 89(1):1–19.

https://doi.org/10.1017/psa.2023.8 Published online by Cambridge University Press