Minds and Machines (2022) 32:1–9
https://doi.org/10.1007/s11023-022-09597-8

EDITORIAL

Minds and Machines Special Issue: Machine Learning: Prediction Without Explanation?

F. J. Boge[1] · P. Grünke[2] · R. Hillerbrand[2]

Published online: 14 March 2022

1 Introduction

Machine Learning (ML) techniques are currently revolutionizing many areas of science. Astrophysicists use them to aid the discovery of rare celestial objects, particle physicists to search for traces of new physics that would otherwise be hard to find, and earth scientists to explore long range correlations, and to accurately predict weather patterns at ever earlier stages. In some cases, however, it is remarkable that the output of the given ML algorithm amounts to a prediction for which scientists as yet see no plausible explanation.

Protein scientists, for example, have struggled for decades to infer the three-dimensional shape of a protein, which largely determines its biological functionality, from the bare sequence of amino acids that it basically consists of. Using traditional modeling, such as modeling based on homologies, this seemed impossible. Yet in the 14th biennial Critical Assessment of protein Structure Prediction (CASP14), a Deep Neural Network (DNN) from Google's DeepMind team, called AlphaFold2, succeeded in predicting structures to within 90% of an average overlap with experimentally determined ones.

For that reason, AlphaFold2 has been called a "game changer" by some (see Callaway, 2020). Yet, others bemoan the fact that it doesn't explain how protein structures arise. For instance, Greg Bowman (2020; emphasis ours), the director of citizen science-based simulation software *Folding@home*, points out that:

> AlphaFold doesn't *explain how* proteins fold, which is another important piece of the protein folding problem. [...] There are also a host of related problems, such as what sort of moving parts do folded proteins have? How do these dynamics enable proteins to transmit information and cargos? How can we

✉ F. J. Boge
fjboge@gmail.com

[1] Interdisciplinary Centre for Science and Technology Studies (IZWT), Wuppertal University, Gaußstr. 20, 42119 Wuppertal, Germany

[2] Institute for Technology Assessment and Systems Analysis (ITAS), Karlsruhe Institute for Technology, Karlstr. 11, 76133 Karlsruhe, Germany

design drugs to turn proteins off (or on)? How can we design proteins to perform new functions?

Hence, there is a sense in which AlphaFold2's remarkable prediction comes *without an explanation*.

This is an important sense, to be sure: AlphaFold2 does not explain how protein folding works. It seems to have somehow learned to bypass the step of explicitly modeling the biological mechanisms leading to the folded protein. Or maybe, there is an image of this mechanism somehow contained in the activation patterns of the nonlinear functions making up AlphaFold2. But this brings us to yet another sense in which the prediction comes without an explanation: AlphaFold2's own functioning in many ways needs an explanation itself.

The implications of this want of explanation in the face of successful predictions are far reaching. For instance, the trustworthiness of ML algorithms in society is a big issue, and it depends, for the most part, on the ability to explain their *functioning*. Ethical issues like these generally profit from philosophers' inputs, as evidenced by the numerous projects and niches on the ethics and societal impact of Artificial Intelligence that we can see originate today.[1]

However, for the sake of using ML in *science*, the other need for explanation—that concerned with an understanding of the mechanisms whose outcome is so successfully predicted by the ML algorithm—certainly obtains a special relevance as well. For, assuming that it remains difficult to understand and explain ML predictions, but that the scientific use of these methods keeps increasing over time, the question arises whether this changes the aims of science from explanation to 'mere' prediction.

These and further issues were first explored, by the editors of this issue, in a workshop organized by P. Grünke and R. Hillerbrand at the Karlsruhe Institute of Technology. This was done as part of a project called *The impact of computer simulations and machine learning on the epistemic status of LHC Data*, in which F. J. Boge is also involved as a postdoctoral researcher. Said project, in turn, is part of an interdisciplinary research unit between physics, philosophy, history and social science, called *The Epistemology of the Large Hadron Collider* and co-funded by the German Research foundation (DFG) and the Austrian Science Fund (FWF).

Much in the spirit of the research unit, the resulting workshop was an interdisciplinary effort, as it involved, next to philosophers, also scholars from the earth sciences (see Boge & Poznic, 2021). Given the fruitfulness of this workshop, the present Special Issue was created as a follow-up publication, even though the contributions to both largely differ.

The essays collected in this Special Issue represent a broad spectrum of perspectives on the issue of explanation in the context of ML, as used in science and beyond. Below, we offer a brief summary of their core theses for the reader's orientation.

[1] Two of us (P. Grünke & R. Hillerbrand) have, for instance, as members of the *AI Ethics Impact Group* (AIEIG), participated in developing a label for AI technologies evaluating them with respect to their ethical acceptability (AIEIG, 2020). Regulations of such technologies are already proposed by the EU and are expected to soon come into effect (European Commission, 2021).

2 Contributions to this Issue

Proceeding by the last names of the (first) author, the first paper by **Falco J. Bargagli Stoffi**, **Gustavo Cevolani**, and **Giorgio Gnecco** investigates the transparency of ML models from the point of view of statistical learning theory and the maxim of Occam's razor; that simpler theories, models and explanations are to be preferred. As the authors rightly observe, simplicity can refer to a lot of things, such as the syntax of the theory, the number of entities postulated, or the metaphysical complexity of the underlying postulates. And clearly, these goals do not always go hand in hand.

The main point of the paper is that, in the context of ML, more complex models can sometimes clearly be preferable for the model's ability to get at the 'truth'; i.e., to offer a representation of the process that generates data of certain, specific types. According to the paper, this holds true when 'complexity' is understood in the sense of the formal learning theory championed by Vapnik (2000). As the reader may or may not know, Vladimir Vapnik, together with Alexey Chervonenkis, introduced a measure for the capacity of a family of ML models, known as the *VC dimension* (see Vapnik, 2000, 70 ff.). For instance, if a perfect binary classifier is understood as a hypersurface separating the data space correctly into two labels (say '+' and '−'), then the VC dimension of a family of such ML models is the greatest number h of data points which can be correctly 'shattered' in this way.

Baragli Stoffi, Cevolani and Gnecco now investigate applications of the VC dimension in the context of supervised ML; i.e., ML with labeled data in the training process. Here, as in other ML applications, there is a well-known problem of balancing the fit of the ML model to the training data with a possible overfitting to that very data; something which is generally tackled by so called *regularization* methods. In particular, one can introduce regularizer-terms into the loss function that determines the learning prescription of the given ML model, and if this is done right, these regularizer-terms can force the model to avoid overfitting.

However, in the context of Structural Risk Minimization (SRM) theory, the regularizer is defined as a function of the size of the data set and the VC dimension, h, of the models used. In this context, h is also sometimes interpreted as a measure of the model-complexity. For example, for linear models and for all models equivalent to linear models, h corresponds to the number of the model's free parameters, which is one well-known way to specify model complexity. Hence, the SRM regularizer may be seen as mediating between model complexity and fit.

These observations are used by Baragli Stoffi, Cevolani and Gnecco to argue that simplicity and truth do not always go hand in hand, as basically suggested by Occam's razor: If the data are sparse, a more complex model may fare better in representing the underlying process that generates the samples of data with their true classes attached. As a corollary, we may note that this yields an interesting perspective on many DNNs in use in practice in science: Since these are often very complex functions in terms of, especially, the numbers of their parameters,

this observation on the relation between complexity and truth may offer one key insight into the reasons for the success witnessed in present-day DNNs, which is not generally well-understood.[2]

The paper by one of us, **Florian J. Boge**, presents a skeptical stance on the possibility of representing, in general, the very data-generating process in terms of ML models such as DNNs. As Boge argues, DNN models should better be seen as continuous with probability models detached from theory, or mere data models; models which are not rich enough in conceptual content (at least on their own) to provide insights into the 'underlying reality' of the data generating process.

This, Boge argues, gives rise to a sense of instrumentality ('c-instrumentality', with 'c' for 'content') which is to be sharply distinguished from the classical notion of instrumentality in models that connects to their want of realistic assumptions ('r-instrumentality'). Furthermore, Boge uses several explainability-studies of ML applications in science that suggest that DNNs can somehow recapture the content of complex concepts, useful also for making predictions, from large data sets without being given direct access to these concepts. Examples are the invariant masses of decayed particles in particle physics, or secondary structures in protein biology.

However, given that DNNs may so rely on information which is not humanly available from the data set, and do so without indicating to the researcher what that information is, they are 'doubly opaque': It is opaque, to a large extent, how they function, but also *what they learn*.

Based on these observations, Boge argues that it is likely that in science using DNNs, there can arise a situation where the DNN makes a genuine discovery but this discovery cannot be explained and understood by the researchers using the DNN. For that, researchers would have to know the relevant concepts learned by the DNN, and thus have to overcome its 'what-opacity'. A specific example where this seems entirely possible is exploratory research using unsupervised ML, which is presently going on, e.g., in particle physics. Thus, a DNN could discover 'paradigm-shifting' new phenomena (to abuse the Kuhnian notion) by relying on yet unknown exotic properties, without making this transparent to the researcher.

Gabe Dupre throws a very different light on explanations in ML,[3] by asking whether opening the black box of successful Natural Language Processing (NLP) DNNs might inform our present theories in Theoretical Linguistics (TL). TL is distinguished by Dupre both from developmental linguistics, which concerns language acquisition, as well as pyscho- and neurolinguistics, which concern either the detailed psychological mechanisms giving rise to linguistic performance or their physiological realization. Instead, Dupre takes TL to be a discipline investigating the

[2] Quite clearly, this is not the final word on ML model performance though. For instance, there is the well-known double descent phenomenon, that the performance of many DNNs first improves, then becomes worse again, and then improves again, as a function of increasing model size, input space size, or training time. For example, Nakkiran et al (2021) argue that the VC dimension is insufficient to explain the double descent of concrete models or training methods, as it only refers to whole families of models.

[3] Unfortunately, the article 10.1007/s11023-021-09571-w by mistake included in other issue but was part of the Special Issue on Machine learning: Prediction Without Explanation? (Dupre 2021).

linguistic competence of already competent speakers at an abstract, computational level, in the sense of Marr (1982).

However, the difference between *performance* and *competence* is crucial here, and Dupre thinks it makes for a principled argument as to why NLP in ML cannot possibly provide insights into TL: Competence concerns a stable, internal capability of processing basic linguistic entities of semantic, morphological, phonological and syntactic guise into utterances that one can observe. The continuity and variation in these utterances is referred to as (linguistic) performance.

The argument that Dupre sees arising from this against NLP's relevance to TL is that a DNN, such as the highly successful GPT-3, will merely learn to reproduce human performance. But empirical linguistics suggests that the relation between competence and performance is highly complex, and so it is unclear, Dupre argues, that NLP systems can learn linguistic competence by being trained on the data left behind by speakers' performance.

A very influential case that has been used to unmask the differences between human cognitive capabilities and those of, in particular, image processing DNNs is discussed by **Timo Freiesleben**: The case of *adversarial examples*. Adversarial examples (short: adversarials) are data instances that, through a small perturbation that, in images, may even be invisible to the human eye, become completely misclassified by an originally successful DNN.

However, while adversarials are generally used to exhibit the problems associated with DNNs, basically the same method of perturbation can also be used to generate counterfactual explanations of the DNN's behavior: Explanations that rely on conditionals of the form "had the input been thus and so, the DNN would have reacted so and so". Thus, given that both data instances that count as adversarials and such instances that give rise to counterfactual explanations are generated by solving the same basic optimization problem, Freiesleben asks what the difference between both really is.

After scrutinizing a number of failed attempts to characterize the relation between adversarials and counterfactual examples, Freiesleben offers his own account, based also on a careful analysis of human interpretability. The dividing line, he argues, really is that adversarials are necessarily misclassified, but not necessarily maximally close in the data space, whereas counterfactual examples are not necessarily misclassified but *are*—quite in line with the Lewisian intuition of evaluating the maximally similar worlds—maximally close in the data space. However, what is still missing, Freiesleben bemoans, is a formally precise notion of *mis*classification.

Concerning the question of how ML may help explanation or even understanding, **Hajo Greif**, in his contribution, paints a quite dire picture. Greif's paper starts to resolve some of the ambiguity in the notion of epistemic opacity that is at the center of many discussions on AI. While many papers on epistemic opacity zoom in on the intransparency on the algorithmic level, Greif argues that what matters is rather the intransparency or transparency of the model. He defines epistemic opacity as a function of the "degree s of an epistemic agent's perceptual or conceptual grasp of a given model" and of "the elements and relations embodied in that model", the model's intelligibility. An important corollary from Greif's account is that opacity or transparency come in degrees.

As Greif continues to argue, the model's intelligibility is not merely or not even primarily a function of the algorithmic complexity and intelligibility, but

depends on the number of (isomorphism) relations between model and target system. This, following Greif, is not essentially different from other kinds of modeling and brings ML in close vicinity with for example computer simulations. Rather than zooming in on ML alone, Greif's paper compares Deep Learning or Deep Neural Network approaches in AI as a subclass of ML with another paradigm of AI, namely Predictive Processing paradigm (Clark, 2013). The latter makes use of connectionist models to explain the functional principles of cortical information processing in humans and other higher animals. While he analyzes the latter as standing in the tradition of analogues models, ML methods are quite generally aligned with digital computer models. As regards the model intelligibility both approaches differ: ML models are argued to be indifferent "towards the endeavor of scientific explanation and understanding" and their role in scientific understanding of explanation is thus significantly undetermined.

Nardi Lam makes a compelling argument that recent accounts in explainable AI that focus almost exclusively on statistical arguments for the validity of AI inferences should be complemented by methods that also study the internal structure of the system. The general concern these approaches in explainable AI address is the very same as this paper: We have a black box system with a high performance of a certain task, but with an unintelligible internal state and want to draw the following inference: Input x led to output y *because* x *has property* p. (Like in: This picture shows a spider because it shows 8 legs) Can we verify that such an explanation is appropriate?

The black-box nature of many AI systems, their opacity to the user, render a statistical analysis not far to seek. Statistical approaches to explainable AI, however, run the risk to reduce ML to yet another, even though more sophisticated method of statistical reasoning with all the accompanying caveats and challenges. Moreover, they also neglect any information on the system's internal state. As for the latter, we do have complete information. Hence it seems that stronger claims than mere statistical ones should be possible. Here Lam takes up and builds on Chomsky's account of knowledge, particularly on the concept of tacit knowledge. As Lam argues in this paper, the implicit use of high-level concepts by ML systems can be seen as an example of tacit knowledge, and hence he argues for the existence of propositional content in ML black box systems. Thereby Lam deploys M. Davies' concept of tacit knowledge that allows to identify tacit knowledge of a rule relating various propositions by a representation of that rule in the systems in the form of a causally systematic process. The question raised above hence can be reformulated in the following way: Is the system aware of p? and Does it use p in order to decide y?

Exploring this, Lam aims to spell out an explanation of ML systems that is not only descriptive of the behavior, but is rooted in the system's computation. These tacit rules bridge the gap between the (lack of) explicit rules in ML systems and the purely descriptive explanations of most explainable AI approaches.

Sanja Sreckovic, **Andrea Berber** and **Nenad Filipovic** analyze key characteristics of ML that make it unsuitable for explanatory purposes. They argue that epistemic opacity and theory-agnostic modeling of ML prohibit access to an explanation of the process through which predictions are reached as well as to an explanation of

the phenomenon which is researched. The main question of the paper is how ML might impact the explanatory practice in science.

The authors thus compare ML models to standard statistical modeling and highlight what is missing in the case of ML models. While statistical models include theoretical assumptions that supplement the data and thereby can offer a potential causal interpretation of the resulting predictions, ML models typically treat the target mechanism as unknown and do not attempt to reflect the causal connections in the target phenomenon.

Sketching the historical relationship between explanation and prediction in science, Sreckovic, Berber and Filipovic show that explanations have value beyond the functional purpose of providing predictions such as understanding, coping with complexity or satisfying curiosity. They claim that ML disrupts the functional relationship between prediction and explanation by providing predictions without explanations. This might lead to different kinds of science, which are explored in two scenarios in the paper, which might develop if scientists search for explanations of ML predictions even if the predictions would be trustworthy without explanations.

The authors predict that ML could lead to a paradigm shift in science, which would diversify science into purely predictively oriented research based on ML-like techniques and, on the other hand, remaining faithful to anthropocentric research focused on the search for explanation.

In an interdisciplinary approach, **David Watson** and **Luciano Floridi**, together with **Limor Gultchin** from the Alan Turing Institute and **Ankur Taly** from Google's Brain team, offer a new view of what it takes to explain something in eXplainable AI (XAI), intended as a general framework that serves to unify several distinct notions present in XAI. The lack of unification in present-day XAI, they claim, is due to the fact that no attention has been paid to the necessity and sufficiency of certain conditions – something they find to be vital to all explanations. However, clearly necessity and sufficiency in the sense of formal logic are too narrow concepts, given the probabilistic nature of much of ML theory. Hence, they build their account around a notion of necessary and sufficient causes, as coined, particularly, by Pearl (2009).

For Pearl, the probability that variable X's taking on value x is a sufficient cause of Y's taking on y is given by $P(y_x \mid x', y')$, where x' and y' are generally different from x and y. That is, it is the *counterfactual* probability that setting X to y would result in X taking on x, given that we in fact observe different values. Analogously, it can be motivated that the probability of x being a necessary cause of y is $P(y'_{x'} \mid x, y)$.

Given these notions, Watson et al. introduce an adaptation to the case of explaining an ML model, f: $\mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X}$ is and input space and $\mathcal{Y}$ an output space. However, for the sake of explaining the pairs (x, y) thus corresponding to f, it is important to also take into account further variables, W, that accompany the inputs $\mathcal{X}$.[4] The 'context' in which f operates is then defined by Watson et al. as the distribution of pairs $z = (x, w)$; i.e., inputs together with non-obvious features that may

[4] We here simplify the notation a bit. Inputs, outputs and contextual variables may of course all be vectors, matrices, or tensors of data.

accompany them. 'Factors', furthermore, are yes/no answers to the question whether a certain z fulfills certain conditions (e.g., x contains the attribute 'female' for gender, and w specifies a lower income threshold).

Armed with this inventory, Watson et al. then extend Pearl's notions in order to define explaining factors for an ML model's decisions, i.e., for a factor c's being necessary or sufficient for outcome $f(z) = y$. Together with a given threshold for the respective probabilities, as well as a partial ordering over the potentially explanatory factors, they then introduce an algorithm they coin Local Explanations via Necessity and Sufficiency (or 'LENS'). Note that this notion of an explanation is sufficiently general to capture, e.g., features visualizable in saliency maps, relevant factors that may be extracted via layer-wise relevance propagation, and many further ideas from XAI that have been suggested as explanatory of an ML algorithm's decisions. The framework is also further explored, in the remainder of the paper, through a series of numerical studies and formal theorems.

Opacity is a major problem for ML. XAI is used mostly for post-hoc analysis of opaque ML models aiming to make them more transparent. **Carlos Zednik** and **Hannes Boelsen** claim that XAI can also be valuable for scientific research and in particular for scientific exploration. Goals of scientific exploration include the identification and refinement of target phenomena, the identification of starting points for future inquiry, and the identification of potential explanations for certain phenomena. Zednik and Boelsen show how XAI methods can be useful to achieve these goals.

Taking up Emily Sullivan's (2019) analysis of the *Deep Patient* model and her discussion about 'link uncertainty', which threatens the scientific utility of ML models, Zednik and Boelsen explicate how XAI techniques for identifying high-responsibility input, such as Shapley Additive Explanation can help to refine the target phenomena and thereby overcome link uncertainty.

Addressing the challenge of efficiently finding good counterfactuals in order to identify causal relationships, Zednik and Boelsen describe the software tool *Counterfactory*, which can produce counterfactual inputs to an ML model that are close to the actual input values yet produce a desired different outcome. This can be used by scientific investigators to create and test new hypotheses about the causal relevance of variables. This might be especially useful in high-dimensional nonlinear systems, in which traditional exploratory techniques often fail.

XAI techniques can benefit the search for explanations in many ways. Zednik and Boelsen highlight their value in cognitive science and argues that they are particularly useful here, because cognitive models often provide algorithmic-level analyses. Zednik and Boelsen argue that it is not surprising that the XAI research program is close to cognitive science as both aim to explain intransparent, complex, high-dimensional systems.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

AIEIG (2020). *From Principles to Practice - An interdisciplinary framework to operationalise AI ethics*. retrieved from https://www.ai-ethics-impact.org/resource/blob/1990526/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb---en-data.pdf

Boge, F. J., & Poznic, M. (2021). Machine learning and the future of scientific explanation. *Journal for General Philosophy of Science, 52*(1), 171–176.

Bowman, Greg (2020). Protein folding and related problems remain unsolved despite AlphaFold's advance, *web-log post*, https://foldingathome.org/2020/12/08/protein-folding-and-related-problems-remain-unsolved-despite-alphafolds-advance/?lng=en

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

Clark, A. (2013). Whatever next? Predictive brains situated agents and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Dupre, G. (2021). (What) can deep learning contribute to theoretical linguistics? *Minds and Machines, 31*, 617–635. https://doi.org/10.1007/s11023-021-09571-w

European Commission (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, COM(2021) 206 final, 21 April 2021, retrieved from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment, 2021*(12), 1240.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.

Sullivan, E. (2019) Understanding from machine learning models. *The British Journal for the Philosophy of Science, axz035.*

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer.