Journal for General Philosophy of Science
https://doi.org/10.1007/s10838-020-09537-z

REPORT

Machine Learning and the Future of Scientific Explanation

Florian J. Boge[1] · **Michael Poznic**[2]

Accepted: 7 November 2020

1 Introduction

On 17 and 18 February 2020, Rafaela Hillerbrand and Paul Grünke (both Karlsruhe Institute of Technology) from the research project "The Impact of Computer Simulations and Machine Learning on the Epistemic Status of LHC Data" organized the workshop "Machine Learning: Prediction Without Explanation?" at Karlsruhe Institute of Technology (KIT).

The project is part of the interdisciplinary, DFG/FWF-funded research unit "The Epistemology of the LHC"; a unique collaboration between philosophers, physicists, historians and sociologists that was recently renewed for three more years.[1]

The workshop's purpose was to bring together philosophers of science and scholars from various fields who study and employ Machine Learning (ML) techniques, in order to create an interdisciplinary setting for discussing the changing face of science in the light of ML's constantly growing use.

Because ML is, from a certain vantage point, nothing but (statistical) optimization executed by digital computers, one may speculate that its increased use exemplifies a paradigmatic turn away from science's traditional aim of explanation, and towards mere pattern recognition and prediction. Moreover, it is also an open question how to explain ML's exceeding utility, as witnessed by various benchmark studies in recent years.

Given this difficult epistemological status of ML, one may ponder the societal implications of its use, as well as its historical and systematic positioning as a scientific method.

Accordingly, the talks' contents will be discussed as organized into (i) practitioners' perspectives, (ii) explanations from ML, (iii) explanations of ML, (iv) societal implications, and (v) global and historical perspectives.

[1] See also http://dailynous.com/2020/01/20/2-6-million-funding-epistemology-large-hadron-collider/.

✉ Florian J. Boge
fjboge@uni-wuppertal.de

Michael Poznic
michael.poznic@kit.edu

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

[2] Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology (KIT), P.O. Box 3640, 76021 Karlsruhe, Germany

Published online: 24 December 2020

2 Practitioners' Perspectives

Among others, Jan Cermak, Uwe Ehret, and Erwin Zehe, three environmental scientists from KIT, were invited to take part in the workshop to share their perspectives on the use of ML in scientific practice. Their talk was organized into three parts, the first one presenting an optimistic perspective, the second a curious, sceptical one, and the third a cautionary, or outright pessimistic perspective.

The idea that the increased use of ML exemplifies a turn from explanation towards successful prediction was not questioned in any of the parts. However, the importance of this was downplayed in the talk by Cermak, who focused on the advantages of ML in reaching reliable results, thus offering an optimistic outlook. One of Cermak's main claims was that ML algorithms in some cases actually outperform the best physics models in environmental science, and act like a well-trained 'sniffer dog'.

This view was countered in part by Ehret, who equally acknowledged ML methods as very efficient tools for data compression with a broad range of applications in science, but, as a partial rejoinder to Cermak's optimistic approach, also formulated desiderata for an interpretable ML, such as 'being right for the right reasons.'

Finally, Zehe, who was one of the invited speakers and chose to include his two colleagues for a more comprehensive perspective, defended the (pessimistic) view that additional environmental modelling will always be needed to understand the causes of climate phenomena. In particular, Zehe defended the view that understanding causation remains a global aim of science, which is not accomplished by ML.

3 Explanation from ML

The question whether it is possible to retrieve explanations from successful ML applications after all was tackled in the talks by Florian Boge (University of Wuppertal) and Thomas Grote (Tübingen University).

In Boge's talk, the example of the Balmer formula and Bohr's atom model was used as an analogy for the kind of gap between scientific understanding and discovery that may be created by the use of ML in science. In essence, Boge's results concurred with those of Zehe: Just as the regularities predicted by the Balmer formula became understandable only against the backdrop of Bohr's model will many ML applications require additional modelling for the sake of understanding their predictions. However, in the case of ML, Boge argued, the gap may become far greater. These considerations were supplanted by case studies on ML's black box nature from particle physics and computer science.

In a similar vein, Grote investigated the intricacies created by ML-applications in medical research. Central to Grote's talk was a distinction between explainability and interpretability, where the former notion was used to refer to the various details that are black boxed in an ML model, the latter to the difficulty of basing justifications on ML results.

As Grote pointed out, the demands for either strongly depend on the aims of the user: Whereas computer scientists, but also medical researchers, are interested in explainability—for the sake of understanding the algorithm itself, or discovering new medical phenomena, respectively—medical practitioners need only care about interpretability. Among other things, Grote sceptically discussed the possibility that multimodal explanations, which offer visualizations as well as natural language explanations, could aid in increasing both.

4 Explanations of ML

The largest number of talks was devoted to explanations of ML. The first of these was that by Tom Sterkenburg (MCMP Munich), which established connections between ML research and formal epistemology. The talk's point of departure was that, since formal epistemology and ML share a common basis, when formal epistemology helps us understand science's success, this may have a carry-over to understanding ML's success.

The main connection Sterkenburg drew was between the problem of induction and 'no free lunch' theorems, which in essence say that no single learning algorithm will perform best across all conceivable tasks. In this respect, Sterkenburg concluded that a lot can be learned for ML explanations from philosopher's approaches to induction.

A similar message was carried by Timo Freiesleben's talk (also MCMP), which demonstrated the use of counterfactual explanations for explainable ML. Freiesleben discussed a formal framework from the ML literature in which the closest possible input is considered that would have resulted in a correct prediction – just as Lewis famously introduced distances between possible worlds to evaluate under what conditions something 'would have been so and so'.

Freiesleben then suggested to relax the assumption that the output space must be interpretable, and to focus on the conditions under which the prediction becomes *in*correct. This could yield an explanation of *adversarial* examples, wherein a small amount of noise is added to, say, an image, which is afterwards completely misclassified by neural networks.

Sergey Titov's (St. Petersburg University) talk drew a connection between explainability in ML and statistical relevance explanations. Titov especially emphasized the role of (homogeneous) partitions in successful statistical explanations. Explaining why something, x, that has attribute A also has attribute B involves finding a partition of the class of all As, such that x's being in cell C_i of the partition makes a difference in probability (i.e., $\mathrm{P}(B|A) \neq \mathrm{P}(B|A \wedge C_i)$).

Based on examples such as an image of a cat being recognized by means of characteristic ears, nose, and mouth, but not in general by the distribution of pixels, Titov then suggested that statistical relevance explanations can be a good model for ML explanations. This conclusion was underscored by a comparison to an explainability framework suggested in the ML community.

Finally, Maël Pégny (Université de Lorraine) distinguished between scientific and pedagogical explanations of ML algorithms. The focus of the talk was on the latter, which are primarily relevant for the lay public. However, Pégny hypothesized that they might be interesting for scientists in several ways as well.

5 Societal Implications

The invited talk by Annette Zimmermann (Princeton University) was the only one to focus explicitly on the societal implications of the fact that ML's functioning is often hard to explain. In particular, Zimmermann dealt with the often cited right to explanations in the context of ML by distinguishing explanation from *justification*.

In the use of opaque ML techniques for decision-making, the apparent lack of explanation might lead to scepticism about the justification of the decisions in question. Her

proposed solution was to seriously consider that explanation may not be necessary for justification.

This was made plausible also with a thought experiment, coined 'accurate Kate': If the fictional character Kate happens to show a strange kind of behavior, in which she spills coffee on people in a cafe apparently at random, but it then turns out that she thereby accurately picks out scammers, we may still demand to know why Kate does what she does, but her behaviour might be considered justified independently of that.

Since ML algorithms have been known to obtain such things as a racial bias through training, Zimmermann concluded that, in policy-making tasks and court judgements, justifying the use of certain algorithms would be crucial independently of their detailed explainability.

6 Global and Historical Perspectives

The two invited talks by Andreas Kaminski (University of Stuttgart) and Johannes Lenhard (Technical University of Kaiserslautern) offered global perspectives on explanation and prediction in the context of ML. Both also used historical evidence to argue for their respective theses.

Kaminski, in his talk "Types of Explanation, Kinds of Reason" started from the question what it means to trust ML models. Based on the assumption that one primarily trusts human beings and that this sort of trust in a non-human entity is something peculiar, he distinguished two types of trust: a normative or ethical one, and an epistemic one.

One intermediate result of the talk was that to trust something like a model means that one can understand or explain this model. With this thesis about explanation, the core theme of the workshop was interpreted in terms of trust: When we look for explanations in the context of ML, we are mainly asking for trust in the mentioned epistemic sense.

Among the most salient points was Kaminski's discussion of Heinz von Foerster's distinction between trivial and non-trivial machines. Non-trivial machines do not easily lead to explanation, understanding and prediction. There are two different kinds of opacity related to the trivial and non-trivial machines, respectively, Kaminski argued. A socio-technical opacity or, black box-ness, related to trivial machines may be lifted, whereas techno-mathematical opacity requires a different type of explanation. The latter often applies to ML and does not straightforwardly allow for simple causal explanations and understanding.

Lenhard's talk, "The History of Mathematization and a New Culture of Prediction", offered a rich historical narrative about the development of mathematical techniques for practical purposes, such as predicting the trajectories of cannonballs. Nicolo Tartaglia's sixteenth century treatise on ballistics, *La Nova Scientia*, here provided a starting point, and Lenhard then outlined how the subsequent debate, which included contributions by Galileo in the seventeenth century as well as Benjamin Robins and Leonhard Euler in the eighteenth century, resulted in the claim that restricting mathematics to tractable forms is constitutive of rationality.

Another example Lenhard discussed was the problem of describing the separation of two components of a mixture in a distillation in thermodynamic engineering. Here, Lenhard detected "an exploratory-iterative culture of prediction": In order to reach better and better predictions, one starts from the ideal gas law, proceeds to the van der Waals equation, which acknowledges molecular details, and further to the Virial equation, which

provides a particularly useful form. These mathematical models are not generally tractable, but simulation modelling makes it possible to use them for predictive purposes.

A core problem here, however, is the presence of free parameters in such equations. Adjusting these results in a proliferation of variants, which constitutes 'the dark side' of simulation modelling, according to Lenhard. Taming this effect requires a balance between theoretical core and adjustable parameters, established in a combination of theoretical and practical considerations. ML methods, however, shift this balance to an extreme, Lenhard claimed, as they involve two important differences to simulation modelling: First, ML channels toward big data, thus marginalizing the theoretical elements, and second, there is to date a mere 'regime', rather than a culture, of prediction in ML.

7 Conclusion

The topic of explanation in the context of ML will certainly remain a topic of interest to philosophers of science in the future. Given the many advances in successful use of ML in science, it may even become a major topic in philosophy of science in the near future. What the workshop certainly established is the fact that different angles are required to approach surrounding questions. Not only are different perspectives on the loss of explanation possible that one might face when ML delivers successful predictions, but one may also seek for explanations of these very predictive successes, root them in the history of science, or investigate their societal implications.

In sum, we may count the strong connection to questions from the history of science, or even to traditional philosophical problems, such as Hume's problem of induction, among the main results, as established in the talks by Kaminski, Lenhard, and Sterkenburg. This was somewhat contrasted by Boge's claim of a profound gap between prediction and explanation, peculiar to ML applications in science, and Grote's emphasis on explainability for the sake of scientific discovery. It remains an open question whether this means a problem for science in practice, as evidenced by the talk of Zehe, Cermak, and Ehret.

In contrast, there are the pressing issues in explaining ML's functioning itself, and especially its failures, as addressed in the talks by Freiesleben, Sterkenburg, Titov, and Pégny. These are most pressing if ML is involved in policy making and political or medical decision making, as established in Zimmermann's talk.

It may also be noted here that a special issue on the subject has been accepted for publication in *Minds and Machines* and is expected to appear in 2021.[2] So the reader interested in getting a deeper view of the issues raised in this report may want to look out for this.

Acknowledgements FJB was employed with the Project "The Impact of Computer Simulations and Machine Learning on the Epistemic Status of LHC Data" of the DFG/FWF-funded research unit "The Epistemology of the LHC" while co-authoring this report, and hence enjoyed funding by the Deutsche Forschungsgemeinschaft (Grant No. FOR 2063). We have also profited from comments by Gregor Schiemann and Helmut Pulte.

Funding Open Access funding enabled and organized by Projekt DEAL.

[2] See https://www.springer.com/journal/11023/updates/18180316.