

# XAI and philosophical work on explanation: A roadmap

Aleks Knoks<sup>1</sup>, Thomas Raleigh<sup>2</sup>

<sup>1</sup>*Department of Computer Science, University of Luxembourg, Maison du Nombre, 6, Avenue de la Fonte, L-4364, Esch-sur-Alzette, Luxembourg*

<sup>2</sup>*Department of Philosophy, University of Luxembourg, Maison des Sciences Humaines, 11, Porte des Sciences, L-4366, Esch-sur-Alzette, Luxembourg*

## Abstract

What Deep Neural Networks (DNNs) can do is impressive, yet they are notoriously opaque. Responding to the worries associated with this opacity, the field of XAI has produced a plethora of methods purporting to *explain* the workings of DNNs. Unsurprisingly, a whole host of questions revolves around the notion of explanation central to this field. This note provides a roadmap of the recent work that tackles these questions from the perspective of philosophical ideas on explanations and models in science.

## Keywords

Deep Neural Networks, Black Box Problem, Explainable Artificial Intelligence, explanation, understanding, scientific models

## 1. Introduction

The last decade has seen an explosion of impressive applications of Deep Neural Networks (DNNs) and other techniques from Machine Learning: systems using these techniques can classify objects from images, diagnose diseases based on medical records, predict protein folds, and do much more. However, these systems are also rightly characterized as *opaque*, meaning that even the engineers of a given DNN can't always understand and explain why it produces a specific output in response to a specific input.<sup>1</sup> This issue – known as the Black Box Problem – has given rise to a flourishing research field of Explainable Artificial Intelligence (XAI) and its range of methods purporting to explain why a given DNN produces a given output, including LIME, SHAP, counterfactual explanations, layerwise relevance propagation, and many others.<sup>2</sup>

The wide variety of existing XAI methods naturally leads one to wonder if some of them are better than others. This, in turn, raises questions of how candidate explanations are to be evaluated: What makes for a correct / good / acceptable explanation? When is one explanation

---

*1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA 2022, University of Udine, Udine, Italy, 2022*


✉ [aleks.knoks@uni.lu](mailto:aleks.knoks@uni.lu) (A. Knoks); [thomas.raleigh@uni.lu](mailto:thomas.raleigh@uni.lu) (T. Raleigh)

🌐 <https://aleksknoks.com> (A. Knoks); <https://thomasraleigh.weebly.com> (T. Raleigh)

🆔 0000-0001-8384-0328 (A. Knoks); 0000-0001-5056-0039 (T. Raleigh)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

<sup>1</sup>For an important early discussion of different forms of opacity, see Burrell [1]. Our focus is on the opacity that Burrell links to the “way algorithms operate at the scale of application”.

<sup>2</sup>We cannot survey these methods for reasons of space. For a detailed survey, see, for instance, Guidotti et al. [2].

better than another? When is an explanation a valuable simplification and when an over-simplification? Since these types of questions have long been pursued by philosophers – especially in the philosophy of science and epistemology – there’s scope for fruitful interaction between philosophy with computer science in the context of XAI. And indeed, several recent publications have applied ideas from contemporary philosophy – more specifically, the literature on scientific models and modelling – to the particular case of opaque DNNs. It pays to distinguish this literature from Miller’s seminal paper [3]. Where Miller called for supplementing XAI with insights from social sciences (within which he includes philosophy) and, in particular, insights about the way “people define, generate, select, evaluate, and present explanation” [3, p. 1], the literature in question debates the viability of the whole project of XAI.

The main goal of this note is to provide a roadmap through this literature. The views we discuss can be arranged on a spectrum. At one end, we have the Optimists who think that the moral to draw from philosophical work on explanation, understanding, and models is that there’s no barrier to the use of XAI models in providing genuine explanation and understanding of DNNs. At the other, we have the Pessimists who argue that there are principled reasons that make these methods suspect, and that there are better alternatives to securing the reliability of DNNs. Our discussion moves from the optimist to the pessimist end (Sections 2–5), concluding with some remarks on promising directions for future research (Section 6).

## 2. The optimist end: Páez and Fleisher

Authors who are most optimistic about XAI are Páez [4] and Fleisher [5]. Páez suggests that the concept of understanding is superior to that of explanation when interpreting DNNs. In fact, he claims that there can be no explanations for black-box systems, at least, not in the traditional sense of the term. This owes to the fact that explanations are *factive*, or that “both explanans and explanandum must be true” [4, p. 445]. Páez argues that XAI models are bound to be false in the same way that an explanation of why a bridged collapsed using Newtonian (as opposed to relativistic) physics is false, since they are only coarse approximations of how the system behaves over a restricted domain. Thus, he writes, “Machine learning is the kind of context in which one can say that, in principle, it is impossible to satisfy the factivity condition” [4, p. 454] on either explaining-why or understanding-why a DNN has produced a specific output for specific input. On the other hand, Páez also thinks that XAI models can provide a non-factive form of “objectual understanding” of a DNN’s internal “mechanisms” equivalent to an engineer’s objectual understanding of a bridge collapsing using a Newtonian model.

Fleisher also focuses on understanding, explicitating it as follows: A subject understands why  $X$  if she grasps an explanation  $E$  of  $X$ , where  $E$  consists in information about the causal patterns relevant to why  $X$  obtains; and where grasping  $E$  means accepting  $E$  and having the abilities to exploit and manipulate the causal pattern information that  $E$  contains. He emphasizes the fact that scientific models are often highly idealized and suggests that XAI methods are relevantly similar. Just like scientific models, such XAI models as LIME represent the causal patterns within a particular DNN and make predictions that are imperfectly faithful to it. In view of this, Fleisher concludes that there is no principled reason for not accepting imperfect XAI models.

### 3. Sullivan

Sullivan's [6] focus is not on the prospects of explanations of opaque DNNs, but rather on whether and when these DNNs themselves – despite their opacity – can lead to explanations and understanding of *target phenomena*. She thinks, for instance, that the melanoma-detection model of Estevan et al. [7] can further one's understanding of mole classification, while the sexual orientation-classification model of Wang and Kasinski [8] cannot provide understanding of the relation between sexual orientation and appearance.

Sullivan suggests that neither the opacity, nor the complexity of DNNs is a barrier to their providing the understanding of the target domain. By extension, she is (at least) not opposed to the use of XAI methods in providing understanding of the way DNNs work. She does think, however, that more is needed for an explanation and understanding of the target phenomenon, namely, evidence supporting the link between the model and the target phenomenon. This conclusion is motivated by the use of models in science – in particular, Schelling's model of segregation [9]. Sullivan argues that a model leads to genuine understanding just in case it provides not a mere “how-possibly”, but a “how-actually” explanation, and that a model provides the latter only if there is evidence that the features of the target phenomenon the model represents do really behave in the way the model has them behave. Thus, Schelling's model is well-situated to provide understanding because there is some (limited) empirical evidence that people's preferences for their neighbours' appearance do indeed cause them to move house. The situation with opaque DNNs is quite different: we do not know what the internal features of a DNN represent about the target phenomenon. Until we have identified the internal representational components within a DNN, it's not even possible to use evidence to reduce what Sullivan calls “link uncertainty”, and so we cannot take the DNN to provide any kind of explanation of its input-output classification. Schelling's model is constructed with clearly labelled representational parts from the start. But with trained DNNs (without XAI) we have no idea which features of the inputs are being grouped together and what sorts of inferences are being performed to reach the output.

### 4. Durán

Durán urges the need for a “top-down” approach to XAI which starts by trying to identify what counts as a bona fide “scientific XAI (sXAI)” rather than adopting a piecemeal ‘bottom-up’ approach of creating a range of purported XAI technologies depending on what computational technology / method happens to be conveniently available. He emphasizes that scientific explanations are meant to grow our understanding of why something is the case, whereas much of contemporary XAI in fact only offers mere classifications and predictions. Durán claims that post-hoc XAI methods are “transparency-conditional”: any explanations, or predictions that the method produces are mediated via the XAI system, rather than engaging directly with the DNN itself. This implies, Durán suggests, that for an XAI model to explain, there must be a formal connection (isomorphism, similarity, or some such) between the DNN and the XAI model. Without such a connection there is no basis for claims that an explanation based on the XAI model applies to the DNN. Durán laments that the form of this connection is never spelled

out. Durán is surely correct here concerning post-hoc model agnostic techniques – and we would add that unless more is said constraining the nature of this connection, isomorphisms and similarities between the XAI and the DNN will simply be too cheap and abundant. Durán also warns that the surveyable and straightforward nature of XAI algorithms make them susceptible to providing a “false sense of explainability because classifications are not explanations” [10, p. 3]. He criticizes the tendency to confuse the “analysis of the structure of explanation” with the “pragmatics of giving explanations”. The fact that different information must be delivered to different audiences has no bearing on the structure of a bona fide explanation. We agree with Durán here and suggest that such a criticism could justly be applied to Miller’s influential paper [3]. Drawing on discussions in philosophy of science, Durán compares the explanations provided by typical post-hoc XAI with an “explanation” of the apparent retrograde motion of planets using the Ptolemaic model of planetary motion. Genuine explanation, Durán points out, is a success term, meaning that it must come with genuine knowledge and understanding of the world. Just as the Ptolemaic model cannot produce knowledge of this kind, neither, according to Durán, can typical post-hoc XAI models produce genuine understanding of the DNN.

## 5. The pessimist end: Durán and Jongsma and Babic et al.

Durán and Jongsma [11] and Babic et al. [12] focus on the applications of DNNs in healthcare and express skepticism about the use of XAI in this context. Durán and Jongsma hold that a typical XAI model doesn’t offer sufficient reason to believe that we can reliably trust the DNN it aims to explicate. On their view, when a layperson sees the appealing visual outputs produced by a post-hoc XAI (such as saliency maps or heatmaps) – she acquires only an *unjustified* belief that it really represents the way the DNN produced the output. The problem is supposed to be that, for all she knows, the post-hoc XAI is as opaque as the original DNN. XAI is said to “induce” the belief that one knows why the DNN produced the output without offering a “genuine reason” to believe that XAI has interpreted the DNN. As an alternative, Durán and Jongsma propose *computational reliabilism*. On this view, one is justified in believing the predictions of a given AI system just in case “there is a reliable process... that yields, most of the time, trustworthy results”. In spelling out the notion of a reliable process, four “reliability indicators” are indicated: verification methods, robustness analysis, a history of (un)successful implementations, and expert knowledge. Jointly these are said to offer a justification to believe that the results of medical AI systems are epistemically trustworthy. Also, such trustworthiness is taken to be necessary, but not sufficient for permissibly acting on an output of a medical AI.

Babic et al. [12] argue against the suggestion that providing XAI should be a legal requirement on using DNNs in a healthcare setting. In their view, XAI outputs are not necessarily the actual reasons behind the outputs of DNNs, nor causally related to them. Babic et al. hold that they provide only “ersatz understanding”, that is, XAI outputs can leave one with a false impression that one understands the working of a given DNN better - see also [13] here. They also criticize post-hoc XAI for failing to be robust, for failing to provide genuine accountability, and for threatening to limit the performance and complexity of DNNs that can be used in healthcare. They conclude that, instead of emphasizing explainability, regulators should focus on ensuring and requiring reliable performance of DNNs.

## 6. Lessons for future research

Having surveyed the literature, we close by identifying two promising directions for future research: (1) and (2). But first, some methodological advice: When reading the literature, it is important to keep in mind the distinction between (i) considering whether an opaque DNN trained to predict or classify phenomena in some target domain might also provide us with explanations of these phenomena, and (ii) considering whether some XAI method can provide us with explanations of the opaque DNN. Some theorists (e.g. Sullivan) are primarily concerned with the former, whilst others (Fleischer, Páez) are concerned with the latter. Often both of these topics will be discussed at different points within a single paper. Furthermore, the term *model* is sometimes used to refer to the full DNN itself (which is said to be a model of the target phenomena) and sometimes to refer to an XAI model of the DNN (thus, a model of a model). The moral here is that this literature doesn't always mean the same thing by *explanation*.

(1) Siding with the Optimists, we tend to think that there is no in principle reason why a simplified, model-agnostic XAI model of a DNN cannot provide (at least some degree of) genuine understanding. The claims made by the Pessimists that there is some kind of fundamental problem with the very idea of such XAI techniques are too strong. However, the Pessimists' core worry that simplifying XAI models may be providing only pseudo-explanations and pseudo-understanding will remain a pressing concern until we have a better grip on when exactly the simplifications / idealizations made by a model are legitimate and useful and when they are not. This is especially challenging in the case of modelling DNNs compared with other examples of scientific modelling. When we employ a simplified model of a physical process or a simplified economic model, we are perfectly aware of how the simplified model is a simplification: we choose what the model represents and which features are being left out of the model, and so we can have a reasonable idea about the importance and relevance of these features. In the case of DNNs we lack an independent grip on the target phenomenon: we don't know how the DNN is transforming the input to obtain the output, and so we can form no clear idea concerning the respects in which a simplified XAI model of a given DNN is a simplification and no reasonable idea as to when the features of the DNN that the XAI model doesn't track might become important. Thus, one crucial topic for future research is to identify some principled basis for deciding when an XAI model is a useful simplification and when it oversimplifies. (And unlike Miller [3], we don't think that the laymen reports could serve as such basis.)

(2) Agreeing with the Pessimists, we think that, at least sometimes, a reliable track record of accuracy should suffice for trusting an opaque DNN. The pressing question, then, is when is such a record enough. In part, this is an ethical issue: when are users / stakeholders owed an explanation for a decision made by a DNN? But there is an epistemological issue here too: under what circumstances is it reasonable to think that future inputs will resemble past inputs, so that past track record of reliability can serve as the basis for trust? How can we estimate variation in future inputs compared with past inputs and the training data? For example, when we think of a DNN trained on a set of standardized photos or scans of one specific organ or anatomical feature, the risk of the system responding in unforeseen ways to new inputs that differ in some crucial way from the training data distribution seems small. But if we think of cases in which the training data and potential inputs allow for more variation, the risk that the system might encounter novel, off-distribution inputs for which it is no longer reliable seems much higher.

One of the four “reliability indicators” Durán and Jongsma identify is robustness analysis – a term taken from engineering, where it refers to an analysis of a systems performance under a range of different conditions. They comment, “Robustness analysis.. allows researchers to learn about the results of a given model, and whether they are an artefact of it (eg, due to a poor idealisation) or whether they are related to core features of the model” [11, p. 332]. This points in the direction of a possible solution to the epistemological issue. However, for now it is also only a promissory note, since it is not immediately clear what a satisfactory robustness analysis of a DNN would amount to, and what sorts of “conditions” would we vary and test.

## Acknowledgments

Knoks benefited from funding of the Luxembourg National Research Fund (FNR) under the OPEN programme within the project Deontic Logic for Epistemic Rights (DELIGHT).

## References

- [1] J. Burrell, How the machine ‘thinks’: Understanding opacity in machine learning algorithms, *Big Data & Society* 3 (2016).
- [2] R. Guidotti, A. Monreale, D. Pedreschi, F. Giannotti, *Principles of Explainable Artificial Intelligence*, Springer International Publishing, 2021, pp. 9–31.
- [3] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38.
- [4] A. Páez, The pragmatic turn in explainable artificial intelligence (XAI), *Minds and Machines* 29 (2019) 441–459.
- [5] W. Fleisher, Understanding, idealization, and explainable AI, *Episteme* (forthcoming).
- [6] E. Sullivan, Understanding from machine learning models, *The British Journal for the Philosophy of Science* 73 (2022).
- [7] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, *nature* 542 (2017) 115–118.
- [8] Y. Wang, M. Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, *Journal of personality and social psychology* 114 (2018) 246–257.
- [9] T. C. Schelling, Dynamic models of segregation, *The Journal of Mathematical Sociology* 1 (1971) 143–186.
- [10] J. M. Durán, Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare, *Artificial Intelligence* 297 (2021).
- [11] J. M. Durán, K. R. Jongsma, Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI, *Journal of Medical Ethics* 47 (2021) 329–335.
- [12] B. Babic, S. Gerke, T. Evgeniou, I. G. Cohen, Beware explanations from ai in health care, *Science* 373 (2021) 284–286.
- [13] Z. Lipton, The mythos of model interpretability, *Queue* 16 (2018) 31–57.