Understanding (and) Machine Learning's Black Box Explanation Problems[*]

Florian J. Boge

Abstract Machine learning (ML) is a major scientific success. Yet, ML models are notoriously considered black boxes, where this black boxness may refer to details of the ML model itself or details concerning its outcomes. Hence, there is a flourishing field of "eXplainable Artificial Intelligence" (XAI), providing means for rendering several aspects of ML more transparent. However, given their tremendous success, why would we even *want to* explain black boxed ML models with XAI? I here suggest that, in order to answer this question, we first need to distinguish between proximate and ultimate aims in using XAI: While the proximate aim may be uniformly to provide instruments for explaining aspects of ML to relevant stakeholders, the ultimate aim varies with the context of deployment. Furthermore, I argue that in science, the ultimate aim is the understanding of scientific phenomena. I then sketch three paths along which understanding of phenomena may be gained by means of ML and XAI. In a coda, I address the possibility of gaining understanding from ML directly, without explanations and XAI.

1 Prelude: Black Boxes, Stakeholders, and the Why of XAI

Machine learning (ML) is a major scientific success, as witnessed by the developments that culminated in two recent Nobel prizes. In the basic problem set-up of ML, we are given a space of data $\boldsymbol{x} \in \mathcal{X}$, where the entries $x_i(\omega)$ of the data instances $\boldsymbol{x}$ characterize the features of some objects $\omega \in \Omega$ of interest. The ML model $M_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$ then maps these data to more informative representations $M_{\boldsymbol{\theta}}(\boldsymbol{x}) = \hat{\boldsymbol{y}} \in \mathcal{Y}$. For instance, the $x_i(\omega)$ could be pixels in an image, representing the local colors of the objects depicted, and $\hat{\boldsymbol{y}}$ could be the class label predicted by $M_{\boldsymbol{\theta}}$ (with the hat indicating that this is the model's prediction, which might diverge from a desired label). Or the $x_i(\omega)$ might be amino acids, and the $\hat{\boldsymbol{y}}$ a protein shape that $M_{\boldsymbol{\theta}}$ predicts for the given sequence $\boldsymbol{x}$. $\boldsymbol{\theta}$ here refers to a range of free parameters of the model that are iteratively adjusted during a 'training phase', so as to make the model perform better and better.

[*] This article has been accepted for publication in the volume *The Role of Artificial Intelligence in Science: Methodological and Epistemological Studies*, edited by Curtis-Trudel, A., Barack, D., and Rowbottom, D. P. and to be published by Routledge.

Prima facie, there is nothing mysterious here: Given enough time and effort, it is usually no problem to look up the detailed algorithm, as well as the model's overall functional form (pending parameter-values), unless the code is proprietary (Burrell, 2016; Rudin, 2019). Nevertheless, ML models are notoriously considered black boxes, and there is a flourishing field of "eXplainable Artificial Intelligence" (XAI), providing means for rendering certain aspects of ML more transparent. What is behind this black boxness?

Many authors (e.g. Beisbart, 2021; Boge, 2022; Burrell, 2016; Creel, 2020; Rudin, 2019) have tried to define the black boxness or opacity of ML, making reference either to the complexity of the function $M_{\boldsymbol{\theta}}$ or uncertainties about how its outcomes arise. However, why would we even *want to* 'explain' black boxed ML models with XAI, given their tremendous success? This question I call 'the why of XAI', and it can only be answered by taking both the 'why' and the 'we' rather seriously.

Thus, consider Humphrey's (2009, 618) seminal definition of epistemic opacity:

> a process is epistemically opaque relative to a cognitive agent *X* at time *t* just in case *X* does not know at *t* all of the epistemically relevant elements of the process.

The process in question may either be the prediction or the learning process of ML (Boge, 2022). But what's crucial for us here is (i) the relativity to an agent, as well as (ii) the mention of epistemically relevant elements. Let us begin with 'epistemic relevance'.

Durán (2018, 108) suggests to interpret 'epistemic relevance' in terms of aspects of ML (or some other model) relevant for the justification of its results. This is plausible insofar as knowledge implies justification on most accounts.[1] However, as Durán and Formanek (2018) point out, justified results might come about without requiring the transparency of $M_{\boldsymbol{\theta}}$: extensive testing, cross-validation, and robustness analysis might be sufficient for justifying the credibility of ML outputs.[2] Hence, justification could in principle be had even without XAI.

However, 'epistemic' must not be read overly narrowly here, as purely referring to knowledge: Many authors believe that opacity has to do with the knowledge *and understanding* an agent has of a model $M_{\boldsymbol{\theta}}$ or its outputs. And since many epistemologists (such as Kvanvig, 2003; Pritchard, 2014) now recognize understanding as the ultimate epistemic good that is ideally delivered with the accumulation of knowledge, it is plausible that those elements of an ML model should also count as epistemically relevant which we need in order to better understand the model. As a corollary, we can see why there is so much interest in the field of XAI, as many authors (such as de Regt, 2017; Khalifa, 2017; Strevens, 2013) trace understanding to explanation.[3] However, for most people outside computer science, understanding a *model* in detail might not actually be of great interest. Thus, circling back: Given their success, for *what sake* do we want to *understand* these models *or* their outputs?

A number of recent proposals have taken the agent in Humphreys' account rather seriously. I will collectively refer to such proposals as 'stakeholder perspectives'. For example, Páez (2019, 441) writes: "the purpose of providing an explanation or an interpretation of a model or a decision is to make it understandable or comprehensible to its stakeholders." Similarly, Langer

[1] Even reliablilism can be read as saying that knowledge is belief that just *is* justified by the reliability of the method of its generation, even though no single agent may be in a position to state that justification explicitly.

[2] For recent accounts of ML robustness, see (Boge et al., MS; Freiesleben and Grote, 2023).

[3] For the connection between XAI 'explanations' and scientific explanations, see (Boge and Mosig, 2025a,b).

et al. (2021, 2) put "conglomerations of stakeholders' interests, goals, expectations, needs, and demands regarding artificial systems" at center stage. Zednik (2021) discusses a concrete example, based on a framework by (Tomsett et al., 2018), wherein different questions about an ML model may be asked by different people, depending on what they want to use the model for. Thus, to answer the why of XAI, we may also need to pay attention to the 'we' in it.

Following the above, different XAI methods might be relevant to different stakeholders, given their different goals and interests: "for instance, LIME is explicitly meant to provide explanations to the users of an ML system while [counterfactual explanations] are meant to do the same for data-subjects." (Buchholz, 2023, 9) To systematically chart the landscape of available methods, Buchholz (2023, 10) suggests to specify that the overall aim of XAI as follows: "Provide instruments that produce explanations of topic *t* for stakeholder *s*", where the "topic can also be spelled out at a more fine-grained level as a particular aspect of an ML method." Hence, might the why of XAI be answered by referring to the aspect to be explained and the stakeholder to explain it to?

I believe we are here back to square one in answering the why of XAI: For, for *what sake* do relevant stakeholders want to explain and understand even *particular aspects* of M_θ or its outputs? In the following, I will propose a novel account for answering the why of XAI, building on a distinction between *proximate* and *ultimate aims*. As I shall argue, Buchholz and others mostly characterize the proximate aims in using XAI, whereas the ultimate aim may differ. Furthermore, as I shall also argue, the ultimate aim varies with deployment context rather than individual stakeholder. Finally, in science, the ultimate aim is the understanding of scientific phenomena, not of ML models.

2 Proximate and Ultimate Aims

Ernst Mayr (1961) famously introduced the proximate/ultimate-distinction into the debate on causes in biology, and he did so by way of an example:

> [A] warbler migrated on the 25th of August because a cold air mass, with northerly winds, passed over our area on that day. [...] the physiological condition of the bird interacting with photoperiodicity and drop in temperature [...] [w]e might call [...] the *proximate causes* of migration. [...] the lack of food during winter and the genetic disposition of the bird [...] are the *ultimate causes*. (Mayr, 1961, 1503)

What I am suggesting here it that, to answer the why of XAI, we first need to similarly distinguish proximate from ultimate *aims*. In the case of the warbler, it might be uncomfortable to rephrase everything in terms of aims: While the warbler may well be said to aim at avoiding the cold by migrating, it is not so clear that he can be said to literally aim for reproduction and survival. However, in the case of ML-stakeholders, no such discomfort should arise: Proximally, they may aim to produce explanations of some particular aspect of an ML model, suitable for understanding these aspects of the model. But ultimately, they might aim for something completely different.

As briefly sketched above, in the remainder of the paper I will argue for the following: (I) While the proximate aim (PA) in deploying XAI may be uniformly specified in the ways suggested by Buchholz, the ultimate aim (UA) is non-uniform – a verdict which is so far compatible with stakeholder perspectives. However, somewhat pace stakeholder perspectives, I will argue that (II) the UA covaries with the *context of deployment*, rather than the individual. In fact, (III) the UA may have *nothing to do with explanation*. Finally, however, (IV) in science, outside the science of ML, the UA is the understanding of the relevant subject matter, *not* the ML model itself.[4]

Since (II) implies (I), and establishing (III) may provide evidence for (II), I will use the next section to establish all these claims by means of examples. The central claim of this paper, (IV) will then be established in sect. 4. In sect. 5, I will sketch three paths along which this aim may be achieved using ML and XAI. Finally, in the coda (sect. 6) I will assess the possibility of gaining understanding from ML directly, without XAI.

3 Ultimate Aims in Medical and Legal Deployment Contexts

The first kind of deployment context I will consider is medical practice (not medical research). In this context, a major issue is the trust invested by patients in what for all we know are relatively reliable methods. For example, as Kundu (2021, 1328) reports:

> a conference posed the following question to its attendees: suppose you have cancer and need surgery to remove the tumor. Which of the two surgeons would you pick if you had to choose between a human surgeon, with a 15% change of dying, or a robot surgeon, with a 2% chance of dying – with the caveat that no one knows how the robot operates and no questions may be asked of it?

Surprisingly, "[a]ll but one of the attendees preferred the human." (Kundu, 2021, 1328) Given the major performance gap between human and machine, stipulated in this example, it is thus understandable that: "Solving the explainability conundrum in AI/ML (XAI) is considered the number one requirement for enabling trustful human-AI teaming in medicine." (Royal Society and Alan Turing Institute, 2019)[1]bienefeld

However, the crucial lesson for us here is the following: In the context of medical practice, the PA may be rendering ML models' workings transparent. A more distal aim served by fulfilling the PA might be making them appear *trustworthy*. But the UA rather appears to be the consentful deployment of the (for all we know) most reliable methods, or even the saving of human lives. XAI clearly serves the PA. But it may thereby also serve the UA by increasing the acceptance of more reliable methods. Notably, the UA thus has nothing to do with explanation, providing evidence for (III).

Take another example, this time involving the legal and ethical implications of decision making. In this connection, Baum et al. (2022, 5–6) ask us to consider the following scenario:

> Suppose [some] company employs a fully automated hiring system to screen, rank, and select job applicants [...] that [...] ranks April in the last place and excludes her

[4] One may also wonder about the aim in deploying ML at all. Often, ML models are deployed in science to make predictions or just to clean and segment data. Here, the aim of using ML might just be to get a prediction or to get clean data, but I suggest that this, too, would be merely a proximate aim. The ultimate aim of using ML within the scientific context would still be to obtain understanding of phenomena, and ML's actual contribution to this might be marginal and mediated by further models and theories. Nevertheless, use of ML would here ultimately serve the aim of understanding phenomena.

> from the further hiring process. [...] maybe this ranking was decisively influenced by the fact that April is a Black woman [...] the worry [arises] that no one can be held morally responsible or legally accountable for excluding April.

However, assume now that a successful XAI method was able to highlight the relevant factors that prompted the hiring system to exclude April from the application process. Would we then not say that an employee of the company's human resources department might be held both responsible and accountable for the (bad) decision, pending their access to the XAI method's output? This is the verdict of Baum et al. (2022, 15): "If [some human decision-maker] had a suitable explanation of the system's recommendation available [...] he [would be] in a position to bear direct responsibility for his decision."

The lesson to be learned here for us, however is again a different one: In the context of moral attribution and legal claims, XAI may serve the PA of making reasons for decisions by ML systems transparent. However, the UA here is the localization of responsibility and accountability. Since an XAI method may also serve the more distal aim of increasing the knowledge had by human decision makers, it might thereby serve the UA of making them accountable.

We see that (I) the UA is non-uniform and (II) varies with the context of deployment: In medical deployment contexts, we ultimately aim to promote reliable methods and save lives. In legal and moral contexts, we aim to attribute responsibility and accountability. In both cases, (III), the UA had nothing to do with explanation. However, why, then, should the UA in science, (IV), be the understanding of the subject matter?

4 XAI and the Aim(s) of Science

Popular accounts of science tell us that "[s]cience aims to explain and understand".[5] This verdict is not just found in popular accounts though, but echoed by several recent accounts within the philosophy of science, most prominently in the debate on scientific understanding (e.g. de Regt, 2017; Elgin, 2017). Not everyone agrees: For instance, Bird (2022, 12) claims that "the aim of science is the production of scientific knowledge". However, insofar as knowledge is valued for its ability to promote understanding (Kvanvig, 2003; Pritchard, 2014), we might consider the production of knowledge a proximate aim of science, with understanding being its ultimate aim.

Furthermore, Rowbottom (2023) thinks that it is wrong to speak of 'the aim of science' in the first place: People have aims, human activities such as science do not. However, Rowbottom (2023, 48) allows to replace talk of the aim of science by talk of 'hypothetically rational aims', where "X is a hypothetically rational aim in doing science if and only if doing science raises the aleatory probability of achieving X more than doing any other possible activity does." Now, several protagonists in the debate on scientific understanding highlight the connection between understanding and abilities (de Regt, 2017; Elgin, 2017; Reutlinger et al., 2018), or even suggest to measure understanding in terms of such abilities. Thus, given the tremendous technological and practical successes human societies have harvested by means of doing science, it seems reasonable to claim that science increases the probability of understanding things much more

[5] https://undsci.berkeley.edu/lessons/pdfs/what_is_science_p4.pdf.

than does any other activity.[6]

The bottom line is that understanding may be quite reasonably held to be ‘the’ aim of science. However, what is the target of the understanding delivered by science? Several authors carefully distinguish between scientists understanding of a given model or theory and their understanding of a given phenomenon or subject matter: de Regt (2017, 23) distinguishes the understanding of a theory, by which he means the ability to use the theory, from the understanding of a phenomenon, which is the aim of science and consists in having an adequate explanation of the phenomenon, based on one’s prior understanding of the relevant theory. Similarly, Strevens (2013, 513; orig. emph.) coins a notion of “*understanding with* [...] a theory”, by which he means being “able to use that theory to explain a range of phenomena”. Of course, understanding with a theory in this way requires understanding that theory itself in the first place (Strevens, 2013, ibid.).

However, recall Buchholz’ suggestion that the aim of XAI is to provide instruments that produce explanations of particular aspects of a given ML method to a particular stakeholder. If this is correct, the target of XAI is a class of models, not a phenomenon out there in reality. How does that square with science’s UA being the understanding of phenomena, or whole ranges thereof?

Before I outline at least three paths along which this may be achieved, let me dispel a distraction: In sect. 1, I followed Durán and Formanek (2018) in arguing that external measures of validation may sometimes be sufficient for establishing reliability, i.e., for justifying the credibility of ML outputs. But sometimes, this may not be enough. For example, Duede (2022) argues that we might want to know whether an ML model obeys certain principles in order to estimate its reliability. Hence, it seems XAI might be used in the service of justification rather than understanding, even within science. Similarly, Scorzato (2024, 17) argues that “lower interpretability makes the assessment of reliability less plausible”, because uncertainty-assessment on the predictions will be harder. Finally, Tamir and Shech (2024, 11) emphasize how “ML research as well as ML applications can suffer from methodological failures calling the validity of inferences based on ML models into question.” In this respect, XAI methods may help to check whether an ML application is actually in line with the envisioned predictive goals.[7]

Do these considerations impair my claim that the UA of using XAI in science is the understanding of phenomena? I do not think so, for even if ML is used mostly for predictive purposes and valued for its reliability – and even if XAI is used mostly for the sake of ensuring this reliability – scientist will ultimately want to understand the targeted phenomena, based on the successful, reliable predictions. Thus, even though ML’s and XAI’s contributions to understanding might be rather indirect, understanding may remain the ultimate aim of deploying ML and XAI in science. Let us see how this might be brought about.

[6] Pace Rowbottom (2023), I am also inclined to think that understanding might come out as the majority vote if one asked scientist about their personal aims in pursuing science.

[7] Cf. also Boge et al. (MS) for an example of how non-empirical considerations may be necessary for ensuring reliability.

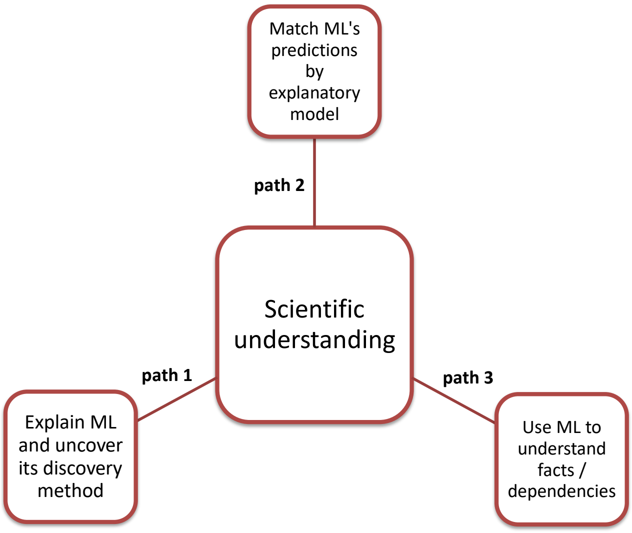

Match ML's predictions by explanatory model
path 2
Scientific understanding
path 1
path 3
Explain ML and uncover its discovery method
Use ML to understand facts / dependencies

Figure 1: Three paths to scientific understanding with (X)AI. Color available online

5 Three Paths to Understanding

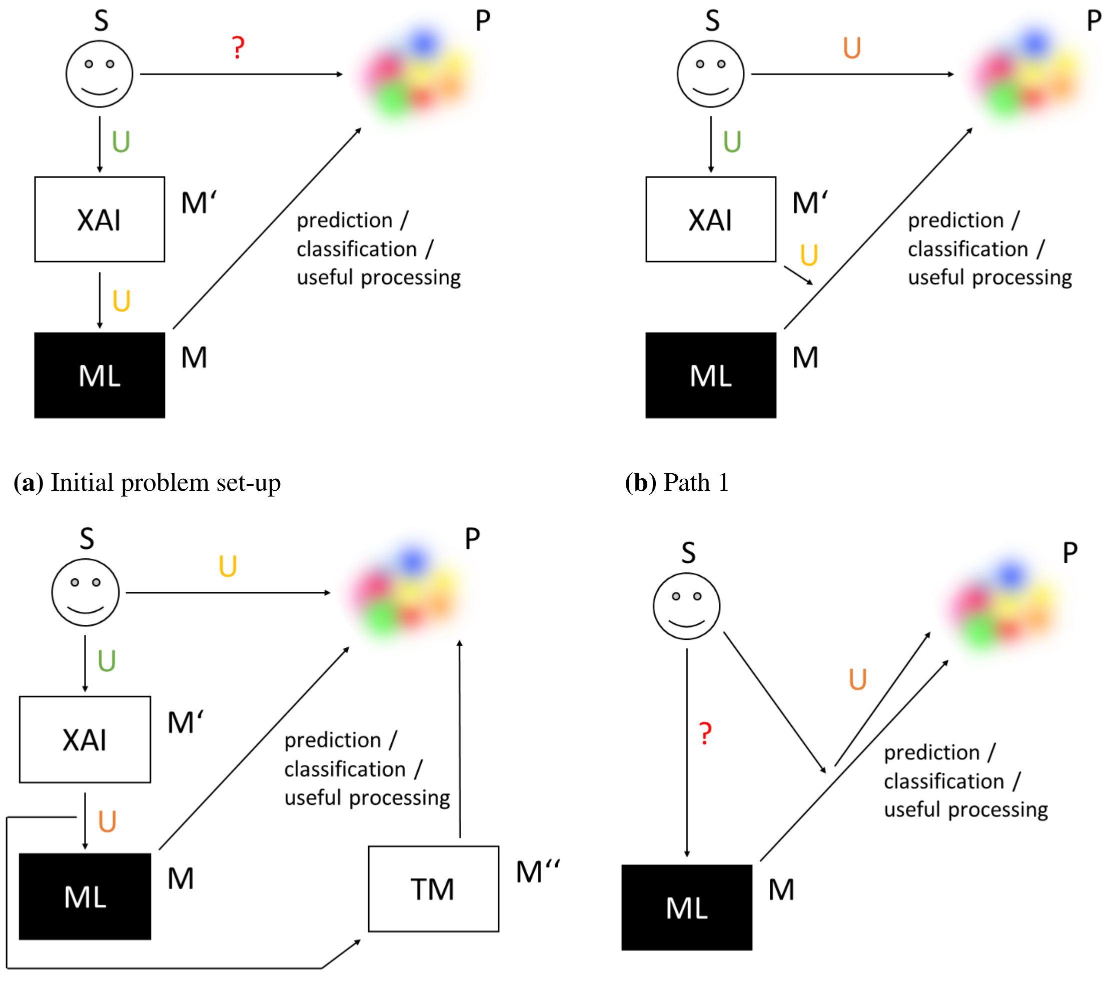
The way I see it, there are at least three paths along which scientist might gain an understanding of phenomena by means of ML and XAI (see fig. 1):

Assume that we are faced with an ML model which has been deployed in a scientific task, such as protein structure prediction, with tremendous success. Then in the first instance, (i) scientists might explain particular aspects of this ML model by means of XAI. This would mean serving the PA of understanding ML, but as a 'side-effect', they might thereby uncover its discovery method and learn a new approach to the subject matter scrutinized with the ML model. This I consider a first path towards scientific understanding from ML, and it clearly proceeds via XAI.

This path towards understanding is fairly indirect though. A more direct one would be to (ii) parallel the successful ML model's predictions by means of a different, explanatory model. Actually, this is not just a possibility, but there are some developments in physics in exactly this direction: Faucett et al. (2021) introduce a method for predicting the presence of a novel signal from physics-inspired variables, by means of a comparison with the decisions made by an ML model. Similarly, Wetzel (2025) utilizes symbolic regression techniques to repackage the information gathered by a ML model into a human-readable symbolic expression.[8] The involvement of XAI here is that it delivers tools to find out what the relevant variables really are, i.e., to uncover *what* the ML model has 'learned' (Boge, 2022).

Finally, (iii), it might be possible to forgo an involvement of XAI altogether and to leverage ML to gain understanding into a phenomenon directly. I will only briefly comment on this possibility in the coda.

[8] A small caveat here is that all of this is being done on well-understood benchmark data, and that it might be significantly harder to do the same thing on novel, poorly understood data (Boge, 2022, 2024).


S
?
P
U
XAI
M'
prediction / classification / useful processing
U
ML
M
S
U
P
U
XAI
M'
U
prediction / classification / useful processing
ML
M
S
U
P
U
XAI
M'
prediction / classification / useful processing
U
ML
M
TM
M''
S
P
U
?
prediction / classification / useful processing
ML
M

(a) Initial problem set-up

(b) Path 1

(c) Path 2

(d) Path 3

Figure 2: Illustration of the three paths to scientific understanding with (X)AI. Color available online

The three paths, along with the initial problem set, are illustrated in fig. 2. In the initial problem set (fig. 2a), a scientist, S, wonders about a fuzzy and colorful phenomenon, P, she doesn't understand (arrow from S to P with red question mark attached). She then uses an ML model, M, to predict or classify certain things about that phenomenon by appeal to data gathered on it (arrow from M to P with description). Using an XAI model, M', which she understands fairly well (arrow from S to M' with green U attached), she may gain some modest amount of understanding regarding the ML model M (arrow from M' to M with yellow U attached), but not necessarily any understanding regarding P.

On path 1 (fig. 2b), she instead directs an XAI model, M', at M's method for prediction (arrow from M' to [arrow from M to P] with yellow U attached). Ideally, in this way, she will not only understand something about M, but also about P, because M's method of access to P will equip S with a novel method for accessing P herself (arrow from S to P, with orange U attached), insofar as she understands that method. For instance, in line with the comments made at the end of the last section, XAI may be proximally put in the service of assessing model reliability. But such reliability-checks too could generate insights about the phenomena under study, for instance, when a model fails in unexpected ways, or when XAI reveals that model is relying on features that scientists hadn't previously recognized as relevant.[9]

On path 2 (fig. 2c), S does direct the XAI model M' at M, but instead of using the understanding so gained (arrow from M' to M with orange U attached) to approach P more directly, she uses it to construct a different model, M'' (arrow from [arrow from M' to M] to M''), which is a theoretical model (TM) that will mirror M's predictions (arrow from M'' to P with description), and may provide explanations which nurture her understanding of P to a fair degree (arrow from S to P with yellow U attached). If successful, it seems likely that a fair amount of understanding of P may be gained in this fashion, which explains the rise of dedicated approaches to this path (Faucett et al., 2021; Udrescu and Tegmark, 2020; Wetzel, 2025).

Finally, on path 3 (fig. 2d), S might forego the effort of even using XAI (arrow from S to M with red question mark attached), and directly inspect M's suggested classifications and predictions (arrow from S to [arrow from M to P]), thereby gaining at least some amount of understanding of P (arrow from [arrow from M to P] to P with orange U attached).

Given that we had argued that the understanding of phenomena is the overall aim of science and that we have thus highlighted three paths along which understanding of phenomena might be gained with ML, two of which involved XAI, it seems clear that 'the' aim of XAI identified by Buchholz (2023) may only be a proximate aim. Further, we also see that in science, the UA of science itself and that of XAI co-align: Within the context of scientific research, both ML and XAI are just further tools that ideally allow us to gain understanding of phenomena, alongside the theories and models that we have already used for this purpose in the past.

6 Coda: Understanding Without Explanation?

I have left open the possibility of foregoing XAI and gaining understanding with ML directly. Why, if this is possible, go the extra mile and use XAI methods? In Boge (2022), I argued that ML models are not in general intended as representations from the outset, and that this distinguishes them from other computational models, especially in their capacity to foster explanations and understanding (Boge, 2019, 2022). Whereas computer simulations contain various elements that can stand in for elements of a targeted systems (however crudely and imprecisely), all that is being represented in ML, at least initially, is a wholesale connection between data $\mathcal{X}$ and representations $\mathcal{Y}$ (class labels, protein shapes...). If I am correct, then this makes understanding directly from ML highly problematic and XAI almost necessary for science's UA (also Räz and Beisbart, 2022).

However, in the preceding section, I indicated an alternative option: On path 3, scientists might forego the use of XAI and use ML for the sake of understanding directly. A very similar account of understanding directly from ML has prominently been offered by Sullivan (2019).

[9] As an anonymous reviewer has pointed this out to me.

Sullivan (2019, 1) argues that “it is not the complexity or black box nature of a model that limits how much understanding the model provides”, but, “a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon” – something she calls ‘link uncertainty’.

While it is certainly right that models need to be appropriately linked to evidence in order to provide understanding, and while some amount of opacity may remain tolerable, it is unclear just how much understanding ML models can promote directly. Thus, Räz and Beisbart (2022, 2) argue that Sullivan’s thesis is only plausible if the understanding gained is “some degree of objectual understanding”.

Objectual understanding is the understanding of a subject matter or phenomenon, P (Dellsén, 2020; Elgin, 2017); as in ‘S understands quantum physics’ or ‘S understands matter interference’. This contrasts with understanding why p, where p is some proposition, such as ‘electrons show interference-behavior’. The two are not identical, and understanding why is usually traced to explanation (de Regt, 2017; Khalifa, 2017; Strevens, 2013), whereas objectual understanding is generally not (Dellsén, 2020; Elgin, 2017).

Elgin (2017) argues that objectual understanding is basic, and in general more encompassing than explanatory understanding; Dellsén (2020) argues that there are cases which can only be objectually understood, but not by means of explanation.[10] It is arguable, though, that we wouldn’t want to call something ‘scientifically understood’ if there was no explanation at all that could be communicated at least among a relevant group of experts (Schuster et al., MS). In any case, it seems right that whatever understanding may be gained along path 3, i.e., from ML directly, will be significantly weaker than understanding gained by means of suitable explanations (Räz and Beisbart, 2022). Nevertheless, embedding this kind of (objectual) understanding into an overall research process, new explanations of related phenomena may be forthcoming (Schuster, MS), and so the existence of path 3 (and with it Sullivan’s (2019) account) remains valuable in its own right.

7 Conclusions

I have argued that, in order to answer the why of XAI – that is, to say why, given their tremendous success, we even want to explain black boxed ML models with XAI – we need to distinguish between proximate and ultimate aims: The proximate aim in using XAI is to provide instruments that produce explanations of particular aspects of ML models for relevant stakeholders. The ultimate aim varies with the deployment context. In science, I argued that the ultimate aim co-aligns with the aim of science itself: The understanding of phenomena. However, this is usually possible only indirectly: by learning something about the ML model’s discovery method, thereby gaining a different kind of access to the subject matter; or by emulating the ML model’s success with a different, explanatory model.

Alternatively, ML may be used directly to understand something about the phenomena in question (Sullivan, 2019). However, following Räz and Beisbart (2022), I argued that this understanding will typically be significantly weaker than the understanding gained by means of explanations. Hence, to answer the why of XAI with a focus on science: Success notwithstanding,

[10] However, see Boge and Stoll (MS) in this connection.

scientists want to explain ML models because this is their best shot at ultimately understanding the phenomena they gather data on.

References

Baum, K., Mantel, S., Schmidt, E., and Speith, T. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy & Technology*, 35(1):12.

Beisbart, C. (2021). Opacity thought through: on the intransparency of computer simulations. *Synthese*, 199:11643–11666. https://doi.org/10.1007/s11229-021-03305-2.

Bird, A. (2022). *Knowing Science*. Oxford, New York: Oxford University Press.

Boge, F. and Mosig, A. (2025a). Causality and scientific explanation of artificial intelligence systems in biomedicine. *Pflügers Archiv-European Journal of Physiology*, 477(4):543–554.

Boge, F. J. (2019). How to infer explanations from computer simulations. *Studies in History and Philosophy of Science Part A*. https://doi.org/10.1016/j.shpsa.2019.12.003.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, pages 1–33. https://doi.org/10.1007/s11023-021-09569-4.

Boge, F. J. (2024). Functional concept proxies and the actually smart hans problem: What's special about deep neural networks in science. *Synthese*, 203(1):16.

Boge, F. J., Krämer, M., and C., Z. (MS). Deep learning for experimental discovery and the theory-freedom-robustness trade-of. unpublished.

Boge, F. J. and Mosig, A. (2025b). Put it to the test: Getting serious about explanation in explainable artificial intelligence. *Minds and Machines*, 35(2):1–28.

Boge, F. J. and Stoll, F. (MS). Extending explanatory relevance. *unpublished*.

Buchholz, O. (2023). A means-end account of explainable artificial intelligence. *Synthese*, 202(2):33.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):1–12.

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

de Regt, H. (2017). *Understanding Scientific Understanding*. Oxford University Press.

Dellsén, F. (2020). Beyond explanation: Understanding as dependency modelling. *The British Journal for the Philosophy of Science*.

Duede, E. (2022). Instruments, agents, and artificial intelligence: novel epistemic categories of reliability. *Synthese*, 200(6):491.

Durán, J. M. (2018). *Computer Simulations in Science and Engineering*. Cham: Springer Nature.

Durán, J. M. and Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, 28:645–666.

Elgin, C. (2017). *True Enough*. The MIT Press. MIT Press.

Faucett, T., Thaler, J., and Whiteson, D. (2021). Mapping machine-learned physics into a human-readable space. *Phys. Rev. D*, 103:036020.

Freiesleben, T. and Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, 202(4):109.

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

Khalifa, K. (2017). *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press.

Kundu, S. (2021). Ai in medicine must be explainable. *Nature Medicine*, 27:1328. https://doi.org/10.1038/s41591-021-01461-z.

Kvanvig, J. L. (2003). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge Studies in Philosophy. Cambridge University Press.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., and Baum, K. (2021). What do we want from explainable artificial intelligence (xai)?–a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473.

Mayr, E. (1961). Cause and effect in biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist. *Science*, 134(3489):1501–1506.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459.

Pritchard, D. (2014). Knowledge and understanding. In Fairweather, A., editor, *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*, pages 315–327. Springer International Publishing, Cham.

Räz, T. and Beisbart, C. (2022). The importance of understanding deep learning. *Erkenntnis*. https://doi.org/10.1007/s10670-022-00605-y.

Reutlinger, A., Hangleiter, D., and Hartmann, S. (2018). Understanding (with) toy models. *The British Journal for the Philosophy of Science*, 69(4):1069–1099.

Rowbottom, D. P. (2023). *Scientific Progress*. Cambridge University Press.

Royal Society and Alan Turing Institute (2019). Discussion paper: The AI revolution in scientific research. https://royalsociety.org/-/media/policy/projects/ai-and-society/AI-revolution-in-science.pdf.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Schuster, A. (MS). Understanding protein folding with deep learning: A case study of alphafold2. *unpublished*.

Schuster, A., Stoll, F., and Boge, F. J. (MS). Objectual or explanatory? distinguishing scientific from scientists' understanding. *unpublished*.

Scorzato, L. (2024). Reliability and interpretability in science and deep learning. *Minds and Machines*, 34(3):27.

Strevens, M. (2013). No understanding without explanation. *Studies in history and philosophy of science Part A*, 44(3):510–515.

Sullivan, E. (2019). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035.

Tamir, M. and Shech, E. (2024). The curve fitting problem, data validation, and inductive generalization in machine learning. *Erkenntnis*, pages 1–15.

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*.

Udrescu, S.-M. and Tegmark, M. (2020). Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631.

Wetzel, S. J. (2025). Closed-form interpretation of neural network classifiers with symbolic gradients. *Machine Learning: Science and Technology*, 6(1):015035.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2):265–288.