Received: 31 May 2021 | Revised: 25 February 2022 | Accepted: 5 April 2022

DOI: 10.1111/phc3.12830

ARTICLE

WILEY

# Philosophy of science at sea: Clarifying the interpretability of machine learning

**Claus Beisbart** | **Tim Räz**

Department of Philosophy, University of Bern, Bern, Switzerland

**Correspondence**
Claus Beisbart, Department of Philosophy, University of Bern, Bern CH-3012, Switzerland.
Email: Claus.Beisbart@philo.unibe.ch

**Funding information**
Swiss National Science Foundation, Grant/Award Number: 197504
Open Access Funding provided by Universitat Bern.

[Correction added on 09-May-2022, after first online publication: CSAL funding statement has been added.]

## Abstract

In computer science, there are efforts to make machine learning more interpretable or explainable, and thus to better understand the underlying models, algorithms, and their behavior. But what exactly is interpretability, and how can it be achieved? Such questions lead into philosophical waters because their answers depend on what explanation and understanding are—and thus on issues that have been central to the philosophy of science. In this paper, we review the recent philosophical literature on interpretability. We propose a systematization in terms of four tasks for philosophers: (i) clarify the notion of interpretability, (ii) explain the value of interpretability, (iii) provide frameworks to think about interpretability, and (iv) explore important features of it to adjust our expectations about it.

## 1 | INTRODUCTION

The recent surge in machine learning is accompanied by concerns about interpretability: Some machine learning algorithms, deep neural networks in particular, are black boxes, their workings are opaque, and their decisions cannot yet be explained or justified. Interpretability has thus become a focus of attention. Very roughly, interpretability is the degree to which we understand, or have insight into, machine learning algorithms. In a recent survey, Biran and Cotton (2017, p. 8) call a system interpretable if its "operations can be understood by a human, either through introspection or through a produced explanation." For Gilpin et al. (2018, p. 81), "[t]he goal of interpretability is to describe the internals of a system in a way that is understandable to humans." The term "explainable" is often used

Beisbart, C., & Räz, T. Both authors contributed equally.

synonymously with "interpretable" (see Gilpin et al., 2018, p. 80) and figures prominently in the buzzword "explainable AI" (XAI, cf. Adadi & Berrada, 2018).

But why exactly is it problematic if machine learning models are opaque? What does interpretability amount to, and how might it be improved? What, exactly, needs to be understood, and what kinds of explanations or insights can enhance this understanding?

These questions lead into philosophical waters, because the clarification of fundamental concepts is a general philosophical task. More specifically, understanding and explanation are central topics in the philosophy of science. Of course, in computer science, "interpretability" has specific, disciplinary connotations, which need not correspond to its meaning in ordinary language or in philosophy. Also, the scope of interpretability – roughly models and algorithms – is narrower than that of explanation or understanding in general. Still, since computer scientists spell out interpretability in terms of understanding and explanation, philosophers can contribute to a clarification of interpretability. It is no surprise, then, that philosophers of science have recently tried to elucidate interpretability.

This paper provides a critical review of the related philosophical work and suggests future lines of research for philosophers of science. Our focus is on machine learning, but our reflections can to a certain degree be extended to, for example, computer simulations. We discuss interpretability and explainability at the same time since neither computer scientists nor philosophers agree on a systematic distinction. We bracket moral concerns about interpretability.

We start with a brief introduction to the computer science background of machine learning in Section 2. After a quick look at strands of philosophical literature that are relevant to a clarification of interpretability, we review the recent philosophical discussion about the interpretability of machine learning in Section 3.

## 2 | MACHINE LEARNING AND ITS INTERPRETABILITY

During the last few years, ML models have been successfully used for many purposes, for example to diagnose skin cancer (Esteva et al., 2017) or to search for exoplanets (Pearson et al., 2018). Outside of science, ML has been employed to build an algorithm that beats human champions in the game of Go (Silver et al., 2016), and to develop self-driving cars (Grigorescu et al., 2020).

Machine learning algorithms (Bishop, 2006; Hastie et al., 2009) stand in contrast to more traditional, rule-based algorithms. Predictions of the latter are fully determined by the creators of the algorithms. In machine learning, by contrast, the creators only specify how the algorithms learn to make predictions. In supervised learning, the dominant ML paradigm, an algorithm learns to predict accurately from data annotated with correct predictions (labels). In unsupervised learning, algorithms detect patterns in unlabeled data. Here we will focus on supervised learning.

Many of the recent successes of ML come from deep learning and are based on deep neural networks (DNNs; Goodfellow et al., 2016; LeCun et al., 2015). DNNs are mathematical models of neurons ordered in a series of several layers. The first layer is the input layer; the input is processed through hidden layers to the output layer, which provides the prediction. DNNs are called deep because of their high number of layers (see Nielsen, 2015).

While DNNs are known to be very successful at making predictions, even computer scientists often do not understand why they are so successful, and what exactly it is that these models learn. And even if scientists know the parameters characterizing the neurons and their connections, they still do not understand the dynamics of the network. This is because the interactions between many neurons are extremely complex, such that the scientists cannot easily map the activity of the neurons to features that we recognize in the data. In this sense, humans do not sufficiently understand these models; DNNs are mysterious black boxes to them (Shwartz-Ziv & Tishby, 2017).

Understanding ML models is taken to be important for various reasons. For instance, when algorithms are used in socially sensitive contexts like criminal justice (Angwin et al., 2016), people subjected to algorithmic decisions arguably have a moral right to understand how the decisions came about. Indeed, the first version of the EU's General

Data Protection Regulation (GDPR, EU, 2016, p. 71) famously grants a right "to obtain an explanation of the decision reached" by an algorithm, and this seems to require some interpretability.

Interpretability has thus become an important research topic in computer science (e.g., Carvalho et al., 2019; Chakraborty et al., 2017; Reyes et al., 2020). Lipton (2018) distinguishes between two broad approaches to interpretability, viz. post-hoc interpretability and interpretability as transparency. Post-hoc interpretability is obtained using a set of techniques that are applied to a trained model and that study the input-output relationship without answering the question of how the model works. For instance, saliency maps highlight those features of an input (e.g., an image) that are most relevant for its classification and thus give the user a sense of what the ML model has been sensitive to. A different post-hoc method (see e.g., Ribeiro et al., 2016) uses simple models that reproduce the predictions of an ML algorithm to some approximation (at least locally). But there are worries that post-hoc explanations of the behavior of current black-box models "are often not reliable, and can be misleading" (Rudin, 2019, p. 206). Scientists who aim at interpretability qua transparency, by contrast, try to better understand *how* a model works (Lipton, 2018). As an aside, we should note that efforts geared toward improving interpretability need not only analyze existing models; researchers can instead try to create new models that are easier to understand (Rudin, 2019).

While there is a lot of technical research on interpretability (and XAI), it lacks a conceptual foundation and is not much integrated with research on understanding or explanation, for example, from philosophy, psychology, and cognitive science (Miller, 2018). In computer science itself, it has been noticed that the notion of interpretability is unclear. Lipton (2018) writes that "few articulate precisely *what* interpretability is or *why* it is important" [emphasis in original]. Murdoch et al. (2019, p. 22071) claim that there is "considerable confusion about the notion of interpretability." Consequently, computer scientist Miller (2018) has turned to philosophy (and to some other "explanation sciences") for help.

Miller's focus is on explanation – a classical topic in philosophy of science (see Friedman, 1974; Hempel & Oppenheim, 1948; Kitcher, 1981; Lewis, 1986; Woodward, 2003, for famous philosophical accounts of explanation; Woodward & Ross, 2021 for an excellent survey). But recently, a new strand of relevant literature has emerged, which is focused on understanding (see Baumberger et al., 2017, for a recent introduction). One reason for this is that many philosophers of science have given up the Hempelian view that understanding is no more than a psychological by-product of explanation (e.g., Hempel, 1965, p. 413). The relationship between understanding and explanation seems indeed less straightforward than some have thought. While one type of understanding, *explanatory understanding*, or understanding *why*, requires explanatory information, a different variety, viz. *objectual understanding* of a domain of things, may not require any explanation (Baumberger, 2011, pp. 70–71; Gijsbers, 2013). Furthermore, for some authors (e.g. de Regt, 2017), even explanatory understanding is more than knowledge of an explanation because it involves certain abilities related to the explanation, but not included in knowledge of it, for example, to anticipate the consequences of a relevant theory without running through a full-blown derivation (but see Grimm, 2006). Work on understanding may thus give a different twist to thinking about interpretability by illuminating aspects of interpretability that are closer related to humans and their abilities.

Yet another strand of philosophical literature related to interpretability is that on the *opacity* of computer-based methods. According to Humphreys (2004, 2009), computer simulations are epistemically opaque because they contain too many epistemically relevant elements – that is, too many computational steps – to be followed by a human in reasonable time (see also Creel, 2020; Durán & Formanek, 2018; Lenhard, 2019; Humphreys, forthcoming). DNNs are opaque in this sense too because the computations done during the training phase as well as those that lead from a specific input to an output in deployment cannot be followed by humans within reasonable time. In a broader sense, the opacity of a method can be understood as its disposition to resist epistemic access, in particular understanding (Beisbart, 2021). This broader sense of opacity can be used to diagnose additional ways in which DNNs are opaque; for instance, we do not know what features they pick up on when they classify images (Boge, 2021; see also Boge & Grünke, forthcoming). To stress the added difficulties of understanding DNNs, Humphreys (forthcoming) argues that they are representationally opaque, that is, they do not represent the target system in a way that allows explicit scrutiny or understanding, since they provide what he calls extensional, implicit and distributed representations. Note

though that some ML methods are *not* opaque in every sense. For instance, humans can understand how decision trees are built up in decision tree learning and how the trees make predictions, at least for small trees.

But how can philosophers use this literature to illuminate interpretability? In the next section, we distinguish between four tasks that philosophers have regarding interpretability and survey existing work on these tasks.

## 3 | INTERPRETABILITY IN PHILOSOPHY

### 3.1 | The meaning of interpretability

A natural starting point for philosophers is to clarify the notion of interpretability using conceptual analysis or explication. Interpretability is, of course, naturally understood as the possibility to be explained or to be interpreted, respectively, where "interpretation" is often tied to understanding. In this vein, Gilpin et al. (2018) require for interpretability that the operations of a system be understandable to humans. But this does not get us very far unless it is clear what explanation, interpretation, and understanding are (cf. Krishnan, 2020; Páez, 2019). Also, we find different definitions of interpretability in the literature; for instance, Murdoch et al. (2019, pp. 22071–22072) require the "extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model." At least prima facie, this is different from understanding the operations, as demanded by Gilpin. These kinds of differences led Lipton (2018, p. 1) to argue that interpretability is not a "monolithic concept, but in fact reflects several distinct ideas." He also stresses that both the motivations for interpretability as well as the means by which people try to achieve it are heterogeneous.

There is room then for conceptual improvement or conceptual engineering. A radical reaction is to give up on the notion of interpretability altogether. Krishnan (2020) goes some way in this direction. Erasmus et al. (2020), by contrast, try to clarify the debate on interpretability by distinguishing the latter from explainability. They argue that explainability is not really an issue, because the outputs of DNNs can be explained according to various accounts of explanation. For instance, a DNN can be regarded as a causal network, which allows for a causal explanation of its behavior. However, such an explanation can be very difficult to understand due to its complexity. For Erasmus et al. (2020), this problem is addressed using interpretation, which is defined as making an existing explanation more understandable by relating it to a second explanation. This leads to a notion of interpretability that is more demanding than that of explainability.

The distinction that Erasmus et al., draw between interpretability and explainability in this way seems rather stipulative to us. The fact that computer scientists strive for explainability shows that it is not straightforward or trivial. Also, the authors' suggestion that interpretation makes an explanation more understandable deviates from common use of the terms. At face value, computer scientists who try to improve interpretability do not have an explanation on the table that they want to interpret. Rather, what they want to understand is an algorithm, a model, and its behavior.

It thus seems more natural to identify explainability and interpretability, to explicate both in terms of understanding, and to make further progress by drawing on philosophical work on understanding. One idea is that interpretability is about objectual understanding (see Páez, 2019, for a similar proposal) because this type of understanding encompasses answers to several types of questions (cf. Zednik, 2019), explanatory questions being just one of them. Explanatory questions may further be diverse and ultimately demand different kinds of explanations. Also, answers to explanatory questions may vary between different methods of ML. Spelling this out in more detail by drawing on philosophical work on explanation may be worthwhile even if, ultimately, no unified picture of interpretability results.

Ultimately, it is not the words "interpretability" or "explainability" that matter. Future philosophical work on interpretability is thus well advised not to limit attention to work that explicitly invokes the labels "interpretability" or "explainability". For instance, in papers that do not use the terms "explainable" or "interpretable", Tishby and Zaslavsky (2015) and Shwartz-Ziv and Tishby (2017) propose to explain the success of DNNs, specifically their good generalization properties, in terms of a theoretical tradeoff between making accurate predictions and finding a

compressed representation (this is called information bottleneck tradeoff). As Räz (2022) has shown, the information bottleneck tradeoff is closely related to statistical explanation, as characterized by philosopher Salmon (1971).

## 3.2 | The value of interpretability

A second task for philosophers is to clarify the value, or, maybe, the role of interpretability: Why is it important to have interpretable ML?

Interpretability may carry intrinsic value, instrumental value, or a combination of both. To establish the *intrinsic* value of interpretability, one may argue that interpretability is intimately tied to understanding, which is often taken to be among the ultimate goals of science. However, this argument does not suffice to establish the intrinsic value of interpretability, because attempts at understanding are typically directed toward phenomena that are found worth studying, and not just any phenomenon.

Perhaps, however, the intrinsic value of understanding suitable phenomena can at least be used to argue for the *instrumental* value of interpretability: Understanding certain phenomena is intrinsically valuable, and often models are means to achieve this goal. Still, to help us reach this goal, the models themselves must be understandable. A simple visualization can help us understand a target system because we can easily grasp it. An opaque ML model, by contrast, may tell us whether a tumor is cancerous, but this does not improve our scientific understanding of tumors, if the model itself is ill-understood. This is the thought behind de Regt and Dieks's (2005) requirement that understanding should be based upon intelligible means (in their case, theories; see also de Regt, 2017).

However, recent contributions have argued that models need not be well understood to provide understanding. Sullivan (2022) claims that it is not our limited understanding of ML models that restricts their power to provide understanding of a target system. Rather, what is limiting understanding is link uncertainty, that is, lack of insight into how the model relates to the target system. One crucial argument for her claim is as follows: Black boxes need not compromise the scientists' abilities to use a model if they merely black-box the implementation of the fulfillment of a known task (e.g. the calculation of a factorial). In a case study about a DNN trained to identify melanoma, Sullivan further argues that scientists' limited knowledge about their opaque models suffices to increase their understanding of melanoma in case link uncertainty is low. A claim analogous to that by Sullivan has been made regarding computer simulations (Kuorikoski, 2011; Sullivan, 2022; Lenhard, 2019; see Jebeile et al., 2021, for a discussion of Sullivan, 2022).

Interpretability may also be instrumentally important for different reasons, for example, because stakeholders that are affected by a decision taken by ML have a right to explanation, and interpretability helps to give the required explanation (EU, 2016). Arguably, however, the right to an explanation targets justification rather than (scientific) explanation: Stakeholders affected by a decision want to know the reasons that justify the decision. Still, an account of the reasons will require some explanation of the decision, so some kind of interpretability is instrumental in securing the right to explanation. Admittedly, though, knowing the reasons does not require a full explanation of how the ML application works (this holds true even though the GDPR grants also the "right to know and obtain communication in particular with regard to [...] the logic involved in any automatic personal data processing", EU, 2016, p. 63; see Corfield, 2010; Schubbach, 2019, for the justification of ML methods and of the decisions; see Vredenburgh, 2022, for an argument in favor of the right to explanation).

Although the two arguments for the instrumental value of interpretability given in the last two paragraphs are to some extent debatable, computer scientists and philosophers agree that interpretability can serve valuable ends. Lipton (2018) mentions five goals that interpretability can serve: to boost trust in a ML model, to support causal inference about the target system, to check whether a model can be transferred to a different range of applications, to obtain additional information about a decision taken by a ML model, and to grant fair and ethical decisions. Similarly, Krishnan (2020) claims that interpretability can help justifying ML models, addressing biases and discrimination against certain groups of people (Angwin et al., 2016), and integrating human judgement and ML models. Zednik and

Boelsen (2020) argue that interpretability, and more specifically XAI, can help determine what a ML model is a model of—in their example, a crucial question is whether a model only traces spurious correlations. According to them, interpretability can also render causal inference possible, and help to produce hypotheses that may facilitate our understanding of human cognition (see also Buckner, 2019 and Sullivan, 2022 forthcoming for this point; see Yoon et al., forthcoming, for the benefits of interpretability in medicine).

Krishnan (2020) adds a different twist to the arguments for the instrumental value of interpretability. She argues that the goals that interpretability serves can be spelled out independently of interpretability, and that the goals should therefore be pursued without reference to interpretability. But we think that this conclusion is too quick. First, interpretability may after all carry intrinsic value. Second, interpretability may be a useful all-purpose means, as is money in daily life. In this regard, it is interesting to note that in a study about XAI, Langer et al. (2021, p. 8) take understanding to be "the pivotal point for all endeavors of satisfying desiderata" set by stakeholders. We thus recommend that philosophers characterize more precisely the ends for which interpretability is useful and the various relations between these ends.

### 3.3 | Frameworks for interpretability

If interpretability covers a plurality of concerns or questions (cf. Section 3.1), philosophers have the task to distinguish different kinds or aspects of interpretability and clarify the relationships between these aspects. In the best case, this leads to a framework for interpretability. Such a framework can be useful to identify strategies how to improve interpretability.

Zednik (2019) proposes a "normative framework for explainable artificial intelligence". It is built upon Humphreys's (2004, 2009) analysis of epistemic opacity and Marr's (1982) *computational levels of analysis* account from cognitive science. Zednik argues that different stakeholders seek different sorts of explanations because they request different kinds of knowledge (in a similar way, Langer et al., 2021 assume that different stakeholders are interested in different desiderata). Marr's account then can help us to locate related explanations at these levels. For example, a creator of an ML model will typically ask questions about how inputs are processed to yield an output (*how* questions), and such questions are appropriately answered at the algorithmic and the implementational levels. By contrast, a subject affected by a decision by a ML model is more likely to ask a *why* question, which Zednik associates with the computational level. Zednik locates various kinds of methods that have been developed to improve interpretability, such as heat maps, in his framework.

Erasmus et al. (2020) use their account of interpretability (see Section 3.1) to propose an alternative framework. As mentioned, their central idea is that interpretation makes a first explanation more understandable by relating a second explanation to it. They thus distinguish several ways in which both explanations may hang together. For instance, they reconstruct the common distinction between global versus local by saying that, in a local interpretation, the second explanation explains only a proper part of the phenomenon explained by the first.

Watson and Floridi (2020) propose a framework with the narrower ambition to conceptualize how we may find the best explanation of one single outcome ("token explanations"). They do this using an explanation game with two agents seeking or giving explanations, respectively. The explanations are at a certain level of abstraction and involve models; the relevant desiderata are accuracy, simplicity, and relevance.

Even if philosophers do not provide a full framework to think about interpretability, they can make more modest, but still useful contributions to clarify aspects of interpretability. In particular, they can assess the prospects of certain kinds of explanations regarding certain explanatory questions about DNNs. In this vein, Buckner (2019) distinguishes between three kinds of explanations for the success of so-called deep convolutional neural networks. An interesting type of explanation that has been discussed is model-based explanation. As Mittelstadt et al. (2019) report, simplified models of ML models pervade much work on interpretability and explainable AI. Although such simplified models can be used for pedagogical purposes and enable users to make predictions about the behavior of the ML model in

question, the authors take the use of simplified models to be problematic because, due to approximations and idealizations, models typically get something wrong (see also Rudin, 2019).

This skeptical attitude towards the power of simplified XAI models contrasts with the fact that many philosophers take models to be powerful devices for understanding, even if the models are simple. In this vein, Páez (2019) is optimistic about the use of comprehensible models for interpretability.

It is, of course, not just philosophers who try to assess the prospects of certain kinds of explanations for interpretability. For instance, physicist Carleo and coauthors (Carleo et al., 2019) point out formal analogies between the extraction of general rules in ML on the one hand and the renormalization group (RG) from statistical physics on the other. In statistical physics, insights into the RG explain certain so-called universalities, so these analogies may help to find explanations of DNN behavior (see Batterman, 2002, for philosophical work on the RG).

All in all, then, it is attractive to systematize the various aspects of interpretability in a framework. So far, several frameworks have been proposed that differ in their theoretical basis and scope, and it is unclear whether they can ultimately be reconciled. The question thus arises how much systematization interpretability allows for.

### 3.4 | Features of interpretability

A fourth task that philosophers have regarding interpretability is to clarify the expectations that one can reasonably have regarding interpretability. Ultimately, this depends on the features that explanations and understanding of ML models have.

A central issue is the extent to which interpretability is *subjective*. Clearly, interpretability would lack interest if it did not spark any kind of insight in individual subjects, but interpretability would hardly be an important topic for research in computer science if it significantly depended on individual feelings or predilections. One aim of Miller's (2018) paper thus is to make research on interpretability independent of computer scientists' intuitions about what counts as good explanation. In a similar vein, Lipton cautions against optimizing interpretability methods to "placate subjective demands" (2018, pp. 21–22). These reservations by computer scientists have a natural counterpart in philosophy, where authors have strived to keep understanding (and explanation) as objective as possible. To this purpose, understanding is decoupled from a subjective sense of understanding, which is not a good indicator of good explanations, but rather strongly influenced by hindsight and overconfidence biases (Trout, 2002). The sense of understanding is not needed if understanding can be explicated using other states or features of a subject, for example, the ability to make certain counterfactual inferences (e.g. Grimm, 2006).

The subjective nature of interpretability often surfaces in claims that interpretability is context- or audience-dependent (e.g. Lipton, 2018; Miller, 2018; see de Regt, 2017; de Regt & Dieks, 2005, for a general contextualist theory of understanding). Different stakeholders ask different questions (Zednik, 2019) and they differ in their expectations regarding explanatory virtues such as accuracy or simplicity (Watson & Floridi, 2020). In our view, a pluralistic outlook that distinguishes between different aspects of interpretability can drive a wedge between the subjective and objective sides of interpretability. It is true that different agents weigh the various aspects of interpretability differently because they are interested in different kinds of questions and prioritize different desiderata. Still, it is a matter of objective fact whether a specific aspect of interpretability has been appropriately addressed.

A feature closely related to objectivity is *factivity*, demanding that an explanation or understanding be based upon facts. So, a second issue is whether interpretability requires factive explanations or understanding. This issue seems to lead into a dilemma. Many DNNs are so complicated that they can only be understood on the basis of simplifications, and, more specifically, idealizations. The latter are often taken to be strategic falsehoods, which violate factivity. Rudin (2019, 207) goes even further and claims that all "explanations [of ML models] must be wrong" because an explanation that was 100% correct would just duplicate the original (complicated) model. But it seems puzzling how false explanations may provide any benefit. The dilemma then is that explanations of DNNs are either too complicated or false, and thus in either case not something that we can use.

In philosophy of science, the factivity of explanation and of understanding are discussed quite generally, and insights from this debate can help us to escape the dilemma. Many (but not all) accounts of explanation require some form of factivity (for instance, in the DN-model of explanation, the premises have to be true). Thus, if one wants to stay true to the idea that we ultimately want explanations, one has to bite the bullet and wait for a time when we know a true story of how DNNs work. Factivity is more controversial regarding understanding. While non-factivists regarding understanding (e.g., Elgin, 2007) stress that simplified models can provide understanding, factivists (e.g., Lawler, 2021) reply that falsehoods are instrumental to understanding, but not really part of its content. Further, as Kvanvig (2009) has shown, if researchers use a model with falsehoods to understand a target system, they need not hold false beliefs; rather, they should use the model with an awareness of the ways in which it goes wrong and thus without taking it to provide the literal truth about the thing to be understood (see Baumberger et al., 2017, Section 4.1 for an overview of the discussion on the factivity of understanding).

One way of resolving the dilemma thus is as follows: We require that explanations of DNNs be factive, but allow that their understanding is not. This allows us to accommodate the reasons that pull in both directions of the dilemma: Actual explanations of how DNNs work are not available yet, and maybe DNNs are too complicated to be explained by us. But this does not preclude some moderate non-factive understanding of them using models. Maybe we can even show that this understanding does not require us to believe falsehoods. Admittedly, though, the last point is controversial; it is false, for instance, if understanding is basically knowing an explanation or getting close to this knowledge (e.g., Khalifa, 2017). Authors who hold such a view can only escape the dilemma by denying one of its horns for both understanding and explanation: Either both do not require factivity, or the factivity of both is compatible with using idealized models. But this is not the place to resolve the general debate on factivity. What we wish to suggest though is that DNNs provide an interesting example to be considered.

## 4 | CONCLUSION

The last few years have seen valuable philosophical work on the interpretability and explainability of ML models. While it is still debated how exactly the terms “interpretability” and “explainability” map to philosophical jargon, it is at least clear that philosophical work on explanation can be brought to bear on the discussion about interpretability. A related philosophical analysis underwrites the idea that interpretability is multi-faceted, first because different kinds of explanatory questions may be raised about ML models, but also because different kinds of explanations may be used to improve interpretability. To systematize the plurality of aspects inherent in interpretability, philosophers have proposed frameworks to think about it. Philosophers can also help clarifying to what degree interpretability is subjective and in which way it is valuable. Although certain arguments for the value of interpretability remain controversial, philosophers agree that interpretability has many benefits.

In line with Hegel's metaphor of Minerva's owl, the largest part of philosophical work on explanation so far has tried to accommodate existing explanations. In the current debate on interpretability, however, things are different: philosophers have a chance to shape ongoing research. To this purpose, they need to address practitioners and not just philosophers of science.

### ACKNOWLEDGEMENTS

We are grateful for extremely helpful criticism by two anonymous referees and for useful comments from the members of the philosophy of science colloquium at the University of Bern. Work on this paper was supported by the Swiss National Science Foundation (grant no. 197504).

Open Access Funding provided by Universitat Bern.

### ORCID

*Claus Beisbart* https://orcid.org/0000-0003-2731-6200

## REFERENCES


Adadi, A., & Berrada, M. (2018). Peeking inside the Black-Box: A survey on explainable Artificial Intelligence (XAI). *IEEE Access*, *6*. https://doi.org/10.1109/ACCESS.2018.2870052

Angwin, J., Larson J., Mattu S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals and it's biased against blacks*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Batterman, R. W. (2002). *The devil in the details. Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press.

Baumberger, C. (2011). Understanding and its relation to knowledge. In C. J. W. Löffler (Ed.), *Epistemology: Contexts, values, disagreement. Papers of the 34th international Wittgenstein Symposium* (pp. 16–18). Austrian Ludwig Wittgenstein Society.

Baumberger, C., Beisbart, C., & Brun, G. (2017). What is understanding? An overview of recent debates in epistemology and philosophy of science. In S. G. C. Baumberger & S. Ammon (Eds.), *Explaining understanding: New perspectives from epistemolgy and philosophy of science* (pp. 1–34). Routledge.

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

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)* (Vol. 8, pp. 8–13).

Bishop, C. M. (2006). *Pattern recognition and machine learning. Information science and statistics*. Springer.

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*. https://doi.org/10.1007/s11023-021-09569-4

Boge, F. J., & Grünke, P. (Forthcoming). Computer simulations, machine learning and the Laplacean demon: Opacity in the case of high energy physics. In A. Kaminski, M. Resch, & P. Gehring (Eds.), *The science and art of simulation II*.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, e12625. https://doi.org/10.1111/phc3.12625

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., & Zdeborová, L. (2019). *Machine learning and the physical sciences. ArXiv:1903.10563v1*. https://arxiv.org/abs/1903.10563v1

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832. https://doi.org/10.3390/electronics8080832

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., & Gurram, P. (2017). Interpretability of deep learning models: A survey of results. In *IEEE SmartWorld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & Big Data computing, internet of people and smart city innovation* (pp. 1–6).

Corfield, D. (2010). Varieties of justification in machine learning. *Minds and Machines*, *20*, 291–301. https://doi.org/10.1007/s11023-010-9191-1

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

de Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press.

de Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*, 133–170.

Durán, J. M., & Formanek, N. (2018). Grounds for trust: Essential epistemic opacity and computational reliabilism. *Minds and Machines*, *28*(4), 645–666. https://doi.org/10.1007/s11023-018-9481-6

Elgin, C. Z. (2007). Understanding and the facts. *Philosophical Studies*, *132*(1), 33–42.

Erasmus, A., Brunet, T. D., & Fisher, E. (2020). What is interpretability? *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00435-2

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115–118. https://doi.org/10.1038/nature21056

EU. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union*, *119*, 1–88. http://data.europa.eu/eli/reg/2016/679/oj

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, *71*(1), 5–19.

Gijsbers, V. (2013). Understanding, explanation, and unification. *Studies in History and Philosophy of Science Part A*, *44/3*, 516–522. https://doi.org/10.1016/j.shpsa.2012.12.003

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th international conference on data science and advanced analytics (DSAA)* (pp. 80–89). arXiv:1806.00069.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*(3), 362–386. https://doi.org/10.1002/rob.21918

Grimm, S. R. (2006). Is understanding a species of knowledge? *British Journal for the Philosophy of Science*, *57*, 515–535.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. *Springer series in statistics* (2nd ed.). Springer.

Hempel, C. G. (1965). Aspects of scientific explanation. In *Aspects of scientific explanation and other essays in the philosophy of science*. The Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*, 167–179.

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

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

Humphreys, P. (Forthcoming). Epistemic opacity and epistemic inaccessibility. In M. Resch, A. Kaminski, & P. Gehring (Eds.), *Epistemic opacity in computer simulations and machine learning*. Springer.

Jebeile, J., Lam, V., & Räz, T. (2021). Understanding climate change with statistical downscaling and machine learning. *Synthese*, *199*, 1877–1897. https://doi.org/10.1007/s11229-020-02865-z

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

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531.

Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, *33*, 487–502. https://doi.org/10.1007/s13347-019-00372-9

Kuorikoski, J. (2011). Simulation and the sense of understanding. In P. Humphreys & C. Imbert (Eds.), *Models, simulations, and representations*. Routledge.

Kvanvig, J. (2009). Response to Critics. In A. Haddock, A. Millar, & D. Pritchard (Eds.), *Epistemic value* (pp. 339–351). Oxford University Press.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & 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. https://doi.org/10.1016/j.artint.2021.103473

Lawler, I. (2021). Scientific understanding and felicitous legitimate falsehoods. *Synthese*, *198*(7), 6859–6887.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Lenhard, J. (2019). *Calculated surprises*. Oxford University Press.

Lewis, D. (1986). Causal explanation. In *Philosophical papers* (Vol. 2). Oxford University Press.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31–57. ArXiv:1606.*03490*.

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

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *FAT* '19: Proceedings of the conference on fairness, accountability, and transparency, January* (pp. 279–288). ArXiv:1811.01439. https://doi.org/10.1145/3287560.3287574

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116

Nielsen, M. A. (2015). *Neural networks and deep learning*. Determination Press.

Páez, A. (2019). The pragmatic turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, *29*, 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pearson, K. A., Palafox, L., & Griffith, C. A. (2018). Searching for exoplanets using artificial intelligence. *Monthly Notices of the Royal Astronomical Society*, *474*(1), 478–491. https://doi.org/10.1093/mnras/stx2761

Räz, T. (2022). Understanding deep learning with statistical relevance. *Philosophy of Science*, *98*(1), 20–41. https://doi.org/10.1017/psa.2021.12

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F., von Tengg-Kobligk, H., Summers, R. M., & Wiest, R. (2020). On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. *Radiology: Artificial Intelligence*, *2*(3). https://doi.org/10.1148/ryai.2020190043

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD '16)* (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salmon, W. C. (1971). Statistical explanation. In *Statistical explanation and statistical relevance* (pp. 29–87). University of Pittsburgh Press.

Schubbach, A. (2019). Judging machines: Philosophical aspects of deep learning. *Synthese*. https://doi.org/10.1007/s11229-019-02167-z

Shwartz-Ziv, R., & Tishby, N. (2017). *Opening the black box of deep neural networks via information*. ArXiv:1703.*00810*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M.,

Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*, 484–489. https://doi.org/10.1038/nature16961

Sullivan, E. (2022). Understanding from machine learning models. *British Journal for the Philosophy of Science*, *73*(1), 109–133. https://doi.org/10.1093/bjps/axz035

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)* (pp. 1–5). IEEE. https://doi.org/10.1109/ITW.2015.7133169

Trout, J. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, *69*, 212–233.

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy*. https://doi.org/10.1111/jopp.12262

Watson, D. S., & Floridi, L. (2020). The explanation game: A formal framework for interpretable machine learning. *Synthese*. https://doi.org/10.1007/s11229-020-02629-9

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J., & Ross, L. (2021). Scientific Explanation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/scientific-explanation/

Yoon, C. H., Torrance, R., & Scheinerman, N. (Forthcoming). Machine learning in medicine: Should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*. https://doi.org/10.1136/medethics-2020-107102

Zednik, C. (2019). Solving the Black Box problem: A normative framework for Explainable Artificial Intelligence. *Philosophy & Technology*. https://doi.org/10.1007/s13347-019-00382-7

Zednik, C., & Boelsen, H. (2020). *The exploratory role of Explainable Artificial Intelligence*. http://philsci-archive.pitt.edu/18005/

## AUTHOR BIOGRAPHIES

**Claus Beisbart** is associate professor for philosophy of science at the University of Bern. He obtained a PhD in physics and a PhD in philosophy from the Ludwig Maximilian University of Munich. His current research focuses on the epistemology of computer simulation and machine learning and on reflective equilibrium.

**Tim Räz** is a postdoctoral researcher at the Department of Philosophy at the University of Bern, Switzerland. In 2013, he obtained his PhD in philosophy from the University of Lausanne. He works on topics like mathematical modeling, scientific and mathematical explanations, and interpretability and fairness of machine learning models.

**How to cite this article:** Beisbart, C., & Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, *17*(6), e12830. https://doi.org/10.1111/phc3.12830