From Explanations to Interpretability and Back

Tim Räz*

March 4, 2024

Written for: *Juan Durán and Giorgia Pozzi (Eds.): Philosophy of Science for Machine Learning: Core Issues, New Perspective, Synthese Library.*

Abstract

This chapter discusses connections between interpretability of machine learning and (scientific and mathematical) explanations, provides novel perspectives on interpretability, and highlights under-explored issues. Interpretability types are proposed: kinds of interpretability should be distinguished using both the parts of ML we want to explain and the parts of ML we use to explain. It is argued that not all explanations are contrastive, and that we should also consider contrasts with respect to models and data, not only with respect to inputs. Theoretical explanations are highlighted; they include issues like generalization, optimization, and expressivity. It is proposed that there are two threats to the objectivity of explanations: One from radical subject-dependence, the other from a lack of factivity. Finally, pluralism is advocated: There are different notions of interpretability and different notions of (scientific and mathematical) explanations. However, the heterogeneity of one area does not transfer to the other in a straightforward manner.

Keywords: interpretability, scientific explanation, mathematical explanation, understanding, XAI, pluralism, machine learning

1 Introduction

This chapter discusses connections between the interpretability of machine learning models from computer science on the one hand, and scientific and mathematical explanation from philosophy of science on the other. Both notions have been discussed for decades in their respective fields, but the relation between them has only been explored in the last couple of years. *Prima facie*, it makes sense to look for connections: Interpretability is concerned with explaining phenomena that arise in machine learning, and it would be useful to know what

*University of Bern, Institute of Philosophy, Länggassstrasse 49a, 3012 Bern, Switzerland. E-mail: tim.raez@posteo.de

such explanations should look like. Also, if explanations of ML phenomena exist, it should be of interest to philosophers of science to determine whether these explanations conform to a mode of explanation they are familiar with, or else to characterize the novel explanation type. Thus, looking for connections can be fruitful for both parties.

This chapter provides novel perspectives on interpretability (see the next section for a working definition) and highlights under-explored issues. One general theme is to steer a middle course between extreme positions on interpretability and explanation. For example, it is argued that while we should not strive for one, homogeneous notion of interpretability, there is hope that we can prevent an excessive proliferation of different notions. Then, interpretability should allow for a certain context- and audience dependence; interpretability has a psychological dimension. But we should not view interpretability as purely subjective and untethered by facts. Finally, the focus on explanations does not mean that the twin notion of understanding is not important (see the chapters on understanding in this volume). Understanding is highly relevant to the notion of explanation, and where it is not, there is room for a complementary role of the two notions.

Here is a section-by-section overview of the chapter. *2. Preliminaries:* I introduce working definitions of the key notions of explanation and interpretability and provide a very brief overview of supervised learning. *3. Types:* I propose to distinguish kinds of interpretability using the parts of ML we want to explain, and also the parts of ML we use to generate the explanation; this leads to interpretability types. *4. Contrast:* Contrastive explanations are useful. However, not all explanations are contrastive. Also, not only contrasts with respect to input values matter, but also contrasts with respect to models, data, and other parts of ML. *5. Context:* A certain extent of context dependence of interpretability is useful or even necessary, but radical context dependence goes too far. *6. Theoretical explanations:* Explanations of general or theoretical phenomena are under-explored. They include issues like generalization, optimization, and expressivity *7. Levels:* If there are explanations on different levels of generality, it is useful to explore how these are related. *8. Objectivity and Idealization:* There are two threats to objectivity: One from radical subject-dependence, the other from a lack of factivity. Both cannot be fully avoided. *9. Pluralism:* There are different notions of interpretability and there are different notions of (scientific and mathematical) explanations. The heterogeneity of one area does not transfer to the other in a straightforward manner.

2 Preliminaries

In this section, I provide working definitions of the notion of explanation as used in philosophy of science (Sec. 2.1), of interpretability as used in computer science (Sec. 2.2), and of the most important aspects of machine learning, supervised learning in particular (Sec. 2.3). These definitions constitute stepping stones for the subsequent discussion.

2.1 Explanation

Scientific explanations have been an important topic in philosophy of science since Carl G. Hempel's (1948) work on the Deductive-Nomological (DN) theory of explanation (see Woodward and Ross 2021 for an overview of scientific explanations). It soon turned out that the DN model (or theory) is not an adequate theory of explanations (Ibid.). Despite efforts during the ensuing decades, no universally agreed-upon theory of scientific explanation has emerged; more on this in Sec. 9. I will use the following working definition. A *scientific explanation* is an answer to a why-question of the form "Why ϕ?", or "Why is it the case that ϕ?" The idea to characterize explanations in this way goes back to van Fraassen (1980); Bromberger (1966). Note that requests for explanations can take other forms, e.g., that of a request of an explanation-how. The entity ϕ to be explained is called *explanandum*, while the answer to the question, the entity doing the explanatory work, is called *explanans*. Take the following, classic example: "Why did the window shatter?" The *explanandum* is the event that the window shatters. A possible answer is: "Because a rock was thrown at it." In this case, the explanation is causal, because the *explanans* consists in citing a cause of the *explanandum*. Not all explanations are causal; there are also structural, statistical, mathematical explanations, which are based on non-causal explanatory relations; examples are given below. For an in-depth discussion of mathematical explanations of ML phenomena see chapter 6 in this volume. Also, explanations may not come in the form of an answer to a why-question; the working definition is a first approximation.

2.2 Interpretability

Interpretability is the problem of understanding properties of ML models, or of classes of ML models, possibly relative to some particular dataset, or to some type of data (see Biran and Cotton 2017; Adadi and Berrada 2018 for computer science surveys and Beisbart and Räz 2022 for a philosophical survey). Research on interpretability encompasses a characterization of what we mean when we say that we want to "understand some aspect of an ML model", a characterization of the properties we want to understand, and the formulation of methods that provide understanding of these properties. Research on interpretability thus encompasses both conceptual problems, such as saying what "understanding a phenomenon" means in general terms, and technical problems, such as formulating methods that provide understanding.

This admittedly vague definition of interpretability does not necessarily capture how the notion is used in computer science. One of the goals of the chapter is to introduce useful distinctions that help to clarify the landscape of interpretability, to distinguish different kinds of interpretability, and to point out connections between different areas of research in computer science that are traditionally not taken to be concerned with interpretability. Note that I will not discuss the notion of explainable AI (XAI) separately from interpretability; rather, questions and methods of XAI are subsumed under interpretability (see

Beisbart and Räz 2022 for a discussion of the relation between interpretability and explainability).

2.3 ML

The focus of the chapter is on supervised learning, one important kind of machine learning (Hastie et al., 2009). The goal of supervised learning is to construct a function F that predicts values $\hat{y}$ (outputs) of variable Y based on values x (inputs) of variable X. Predictions are written with a hat, $\hat{y} \in Y$, and ground truths, which are part of the dataset, without a hat, $y \in Y$. In order to construct such a model, a dataset $D = \{(x_i, y_i), i \in 1, ..., n\}$ is used. D is sampled from a system in the world S, such that the $x_i \in X$ are instances of the variable for which we want predictions, and $y_i \in Y$ are instances of the variable to be predicted. For example, the x_i could be images of different animals, y_i would be correct labels of the animals depicted ($y_i \in$ {cat, dog,...}) and the problem would be to predict which animal is depicted in a given image.

The idea is to let a model M learn the function $F : X \to Y$, thus approximating the relation between X and Y encoded in D. More specifically, M constructs an F that minimizes the error on instances in D according to some loss function. The function F can be seen as a property (the input-output profile) of the model M. The model M "learns" F through some optimization procedure O. Learning means that the parameters of M (model parameters) are adapted such that the distance between F and D is minimized as much as possible. In order to check whether M has succeeded in approximating the data, the dataset D is usually split into a training and a test set. The training set is used to learn the function F; the test set is used to check whether the learning was successful, that is, whether F actually approximates D on samples not used in constructing F. An ML model M is successful relative to a dataset D if it has a high accuracy (small loss) on the test set. To return to an example, an ML model has successfully learned to predict (or classify) animals if it is able to predict, with high accuracy, animals depicted in images *the model has not seen during training*. A model that succeeds in doing this is said to generalize well. The most important aspects of supervised learning are summarized in figure 1.

$$S \overset{sample}{\rightsquigarrow} D \overset{approx}{\longleftarrow} F \overset{compute}{\longleftarrow} M \overset{learn}{\longleftarrow} O$$

Figure 1: Components of supervised learning. Legend: S – System to be modeled; D – dataset with structure $\{(x_i, y_i), i \in I\}$ sampled from S; $F : X \to Y$ – function approximating the relation D; M – ML model computing F; O – optimization procedure for M to approximate F.

This overview is very general and incomplete. More can be said about which model is appropriate for which task. For example, while deep neural networks (DNNs) may be appropriate for unstructured, high-dimensional data, other models (random forest, boosted trees) may be more appropriate for tab-

ular data, where individual variables have an "intuitive meaning" (see LeCun et al. 2015; Goodfellow et al. 2016; Hastie et al. 2009 for details). The same ist true for finer points concerning architecture, hyperparameter tuning, optimization procedures, and so on. Then, supervised learning is not the only learning paradigm; there are also unsupervised learning, reinforcement learning, and generative modeling. Also, classical supervised learning may no longer be the most successful paradigm in view of self-supervised large language models. Finally, the situation as depicted in the figure lacks context. ML, and interpretability, are affected by issues such as: What kind or class of ML model is used in what empirical context? What is the goal of the use of an ML model (use in natural science, public administration, ...)? Who is affected by model output (directly affected decision subjects, indirectly affected by consequences of ML models, ...)? The discussion below will fill in some of these aspects.

3 Types

How are interpretability and explanations related? Simply put, there are different ML phenomena we want to understand, and understanding may be provided by explanations of these phenomena (see Lipton 2018; Zednik 2021; Creel 2020). We have just seen (Fig. 1) that supervised learning has different parts. A simple consequence of this is that different explanations may be appropriate for phenomena from different parts of ML, and that different explanations correspond to different kinds of interpretability. How should different kinds of interpretability be classified or distinguished? The simplest way is to distinguish them by the part of ML they belong to. For example, a well-known distinction (Lipton, 2018) is between post-hoc interpretability, which concerns properties of the predictor function F, while transparency concerns properties of the model M, e.g., its parameters.

In this section, I propose a simple refinement of this well-known idea. The refinement is based on a systematic use of the *explanandum-explanans* distinction. The idea is to distinguish kinds of interpretability by specifying not only their *explanandum* type, but also their *explanans* type. Recall that the *explanandum* is the property or phenomenon we want to explain, while the *explanans* is what is doing the explanatory work. To specify the *explanandum* and *explanans* type, we state which ML part they belong to; see figure 1 for the parts. If we write the explanatory relation between *explanandum* and *explanans* as $\rightarrowtail$, we can specify an *interpretability type* in the following form:

$$\text{explanans type} \rightarrowtail \text{explanandum type}. \tag{1}$$

The idea of specifying an interpretability type in this way is borrowed from mathematics. To define a mathematical function, one first specifies a domain and a co-domain (target). For example, a metric on X is a function $d : X \times X \rightarrow \mathbb{R}_{\geq 0}$, such that ... . The domain and target do not (in general) determine the function, but they tell us where it lives. This, in turn, tells us about certain properties of the function. For example, the domain of a metric makes sense

because we want to examine how close pairs of elements of X are, and we want closeness to be measured on a scale of non-negative real numbers. Functions not defined on this kind of domain and target cannot be metrics.

Here is how this translates to interpretability. Assume that we want to explain a particular output $\hat{y} \in Y$ on the basis of a given, trained model M, but we do not want to take the optimization procedure into account. Then the explanation we are looking for has interpretability type $M \rightarrowtail \hat{y}$, where $\hat{y} \in Y$. Note that this is the usual setting of many XAI methods, such as saliency maps. If we want to explain properties of the predictor function in terms of the model and the optimization procedure, then we are looking for an explanation of the type $\{M, O\} \rightarrowtail F$. This is the setting of the so-called Information Bottleneck (IB) method, which examines learning phases of models to illuminate the generalization properties of DNNs (see Shwartz-Ziv and Tishby 2017, and Räz 2022 for a philosophical discussion).

Why is it useful to distinguish interpretability types with relation (1)? First, we can capture commonly discussed kinds of interpretability. For example, the distinction between understanding *of* an ML model and understanding *with* an ML model can be rephrased as: if we are after (explanatory) understanding *of* the model, the *explanandum type* will be in the set $\{F, M, O\}$. Understanding *with* the model, on the other hand, would be provided by an *explanandum type* in the set {D, S}, while the *explanans type* should include one of $\{F, M, O\}$; otherwise, it would be an explanation without ML. Second, and more importantly, interpretability types can help us draw finer distinctions and clarify misunderstandings. As mentioned above, the distinction between *transparency* and *post-hoc explanations* was introduced by Lipton (2018). On the basis of interpretability types, we can refine this idea as follows. We may be in a situation in which we have access to the structure of a model (it is transparent), but we do not know how it was trained. In this situation, all we can hope to provide is a *post-optimization* explanation. This is an interpretability type such that O is neither in the *explanandum* nor the *explanans*. Usual XAI methods, which try to explain outputs, are of this type, but also methods that try to align inner parts of a model with human-interpretable concepts, such as TCAV (see Kim et al. 2018 on TCAV and Räz 2023 for a philosophical discussion). In a different situation, we may only have black-box access to a model, that is, we can feed it inputs and get outputs, but we do not know how information is processed. In this situation, one could try to feed F with novel data D' and try to understand its behavior on D'. Such a *black-box* explanation would be of type $\{D', F\} \rightarrowtail F$. In this case, the interpretability type should not contain M or O. Finally, we may be in a situation where we are interested in model behavior and we have full access to both M and O. Explanations of type $\{M, O\} \rightarrowtail F$ are not concerned with transparency insofar as we do not want to understand or explain the inner working of a model. It may still be useful, or even necessary, to use M and O as part of the *explanans*, and transparency is required for this interpretability type.

This last scenario is relevant for practical reasons. It is sometimes claimed that certain interest groups (stakeholders) such as decision subjects and policy

makers are not interested in the inner workings of ML models. Using interpretability types, one recognizes that this statement is ambiguous and potentially misleading. It may be true that decision subjects primarily want to understand how predictions came about, which means that they are not interested in M and O as *explananda*. However, they may still be interested in explanations of model outputs that draw on M and O, that is, in the interpretability type $\{M, O\} \rightarrowtail F$. Such explanations may provide insights that cannot be gained by a black-box method. Thus, it is a misconception that information about the inner workings of ML models are irrelevant for decision subjects just because their requests for explanations may not concern these aspects.

In the above, I have been rather loose in describing interpretability in terms of both explanations and understanding. While there are differences between these concepts, at least as discussed in philosophy of science, the concept of interpretability types may be fruitful when discussing different kinds of understanding as well: we simply have to distinguish the object we want to understand (in parallel to the *explanandum*), and the means, or information, that we want to draw on to gain understanding (in parallel to the *explanans*). The importance of understanding ML is also emphasized in chapter 9 in this volume.

4 Contrast

Distinguishing kinds of interpretability by type is a useful classificatory device, but it does not uniquely determine all kinds of interpretability. More fine-grained distinctions can be drawn by using contrast (see Miller 2018, Sec. 2.3.; Guidotti 2022). In philosophy, the idea that explanations are contrastive was probably first emphasized in van Fraassen's (1980) pragmatic theory of explanation. Discussions of contrast are part of a wider discussion about the extent to which explanations are context dependent; more on this below. Let us start with an example (adapted from Skow 2016):

(Q1) Why did Adam eat the apple?

(Q1) is a reasonable request for an explanation, but its scope, the kind of answer one might be after by asking it, is very wide, and depending on the situation, different answers may be appropriate. For example, the question can be read with the following emphasis:

(Q2) Why did *Adam* eat the apple?

(Q2) focuses on Adam's relation to the event in question, and an appropriate explanation focuses on how Adam, as opposed to someone else, ended up eating the apple. An appropriate explanation could be: "There was no one else there." The question can also be read with a different emphasis:

(Q3) Why did Adam eat the *apple*?

(Q3) focuses on the apple's relation to the event, and an appropriate explanation focuses on the choice of apple as opposed to some other food. An appropriate explanation could be: "There was pie there, but Adam is watching his diet." These two examples suggest that the role of contrast is to constrain the kind of explanatory information we are looking for: (Q2) and (Q3) restrict the appropriate answer to certain parts of the causal past of the *explanandum* phenomenon, whereas the original question (Q1) does not provide such a restriction.

Now consider some analogous examples in the context of ML. An important case is the explanation of a particular outcome $\hat{y}$ of an ML model M:

(Q4) Why did decision subject i get prediction $\hat{y}_i$?

(Q4) is a request for explanation with no (explicit) contrast. Like (Q1), this question has a wide scope – in principle, it concerns the entire history leading to the prediction $\hat{y}_i$. It is possible to formulate more specific versions of (Q4), which highlight what kind of explanatory information we seek. For example, assume that F is a binary predictor, and that decision subject i has received a decision $F(x_i) = \hat{y}_i = 0$ based on input x_i. Decision subject i may now wonder what it would take to get a different decision $\hat{y}_i = 1$ from the same predictor F. This kind of explanation has been discussed as "counterfactual explanations" of algorithmic output, starting with Wachter et al. (2018) (see Buijsman 2022; Zerilli 2022 and chapter 6 in this volume for discussions of counterfactual and manipulationist notions of explanations in AI). We can thus reformulate (Q4) as:

(Q5) Why did i get prediction $F(x_i) = \hat{y}_i = 0$, as opposed to getting prediction 1 from F?

(Q5) can be read as a question about the general properties of F that yield predictions 0 and 1. Now, (Q5) may still be too general if decision subject i is only interested in particular contrasts, for example, in inputs that are close to their input x_i according to some metric. This interest may be motivated by the desire to determine which changes to x_i would flip the decision and require the least effort. In this situation, (Q5) specializes to:

(Q6) Which values x are close to x_i (according to some metric), but such that $F(x) = 1$?

(Q6) is an even more focused request for information about the predictor F, in that information about F "far away" from the value x_i is considered to be irrelevant. Thus, (Q6) excludes a lot of information that would be relevant in response to (Q4). And, despite not being a why-question, it can be interpreted as explanatory because, while it is more focused than (Q5), it requests information about properties relevant to classification, in particular, information about the ball around x_i in input space that gets the same prediction as x_i.

Are all explanations contrastive, as suggested by Miller (2018)? Not necessarily. A first example is (Q4), the open-ended request for an explanation of a

particular output. This is not a contrastive request for an explanation. It may be objected that while (Q4) is not contrastive, it is not possible to fully answer (Q4), because it is too open ended. However, if the request only concerns information about the predictor F, and if F is sufficiently simple (e.g. a linear model of few variables), one can provide a full answer by providing a description of F.

A second example of a non-contrastive request requires very specific information. A decision subject asking (Q4) may be interested in the input x_i that led to the prediction $F(x_i) = \hat{y}_i$:

(Q7) Which input x_i yielded the prediction $F(x_i) = \hat{y}_i$ for decision subject i?

This is not a contrastive request, and it is also not a why-question, which means that, on the surface level, (Q7) is not a request for an explanation, as opposed to (Q4). The reason why I would nevertheless suggest that (Q7) is a request for a (partial) explanation is that (Q7) asks for a specific fact that partially determines, or is responsible for, the output, which is the phenomenon we want to understand. It therefore constitutes a request for explanatorily relevant information. Also, (Q7) is a special case of (Q4), and the answer, the value x_i, does provide a (partial) explanation of the prediction. Note that accepting (Q7) as a request for an explanation means that we do not strictly adhere to the working definition of what constitutes an explanation.

Insisting that all explanations are contrastive may also be problematic because contrasts narrow down the scope of acceptable explanations. Presupposing a particular contrast leads to the exclusion of certain answers. Let us consider a different version of (Q4) to see why this is problematic. One legitimate concern one may have when asking (Q4) is with properties of the decision process, such as the predictor function F or the model M used to obtain the decision $\hat{y}_i$. One may ask:

(Q8) What is the predictor function F (or model M) that yielded the prediction $\hat{y}_i$?

An answer to this question may give rise to a contrastive follow-up: Once one knows F, it may be asked why F (or a certain class of predictors) was used as opposed to a predictor F' (or a different class of predictors), which may have yielded a different prediction. To give an example, in certain contexts, black-box models may have no advantage in performance over interpretable models, but their decision process is inscrutable; this is proposed by Rudin (2019). This alone may constitute sufficient grounds to challenge decisions reached with black-box models. Compare this to the contrast in (Q6), which focuses on inputs with respect to the *same* predictor F. By narrowing down the explanatory request, this question necessarily excludes other aspects of ML, such as the choice of model. However, the aspects that are excluded by a contrast may be exactly those that need to be explained. Instead of asking how the decision subject should change in view of a fixed predictor, as suggested by the contrast in (Q6), it may be more appropriate to ask whether a different model should be used, as suggested by (Q8). In principle, there is no limit on the aspects of ML that can

and should be scrutinized by asking explanatory questions about them, from the predictor F, to the model M, to the optimization, data, or even the very fact that decision are made at all. Thus, an open-ended question, while unfocused, encompasses more possible contrasts with respect to an *explanandum* and may thus be preferable for some purposes.

5 Context

To what extent are explanations context dependent? Context dependence means that both the interpretation of an explanatory request, and what constitutes an appropriate answer to this request, depend on the situation in which the request for explanation is posed. In particular, what the explanatory request is and what constitutes an appropriate answer, may depend on a) the kind of ML model in question, b) the empirical domain or problem to which the model is applied, c) the audience asking the question or receiving the explanation, and d) the purpose behind the request for explanation, and possibly further factors. The context dependence of scientific explanations was discussed extensively in the wake of van Fraassen's (1980) pragmatic theory; see also Woodward and Ross (2021, Sec. 6).

Context dependence is an important feature of explanations, but it also creates challenges. For example, it may lead to a proliferation of explanations, it can make it hard to specify what counts as a good explanation, or to provide a "rigorous definition" (Doshi-Velez and Kim, 2017) of explanations. Context dependence also threatens to trivialize explanations, say, if the person requesting the explanation can decide freely whether a given answer is adequate or not. The challenge is to allow for some context dependence, while not pushing contextualization too far by letting what constitutes an appropriate explanation depend on, say, the mood of the person asking for it. The idea that there may be two kinds of context dependence, one that is mostly harmless, and one that is more radical, is articulated in Woodward and Ross (2021). Two reasonable kinds of context dependence have been discussed above. First, if we specify the source and the target domain of an explanation (Sec. 3), this may yield several, distinct classes of explanations. Interpretability types place restrictions on both the model in question, the empirical domain (if any) about which the question is asked, and the resources that can be used to answer the question. However, the type does not determine the explanation. The second kind of context dependence is contrast (Sec. 4). A precise articulation of the explanatory question narrows down what kind of explanatory information is considered relevant. This may yield different kinds of explanations. Also, by narrowing down what kind of information is considered relevant, contrasts may help with audience dependence, because different kinds of information are relevant for different audiences.

Some kinds of context dependence are not captured by contrast and interpretability types. For one, there are more fine grained kinds of audience dependence; see Langer et al. (2021) for a stakeholder-centric perspective on

XAI, and Zednik (2021) for a proposal of which parts of ML different stakeholders are interested in. For example, experts and laypeople may have the very same explanatory request, with respect to the same explanation type, and still have different explanatory needs: An explanation that is comprehensible to an expert may be inaccessible to a layperson. This sort of context dependence is important, but it has traditionally not been central to the discussion of scientific explanations in philosophy of science, because the subject matter of scientific explanations are the best explanations that science has to offer to experts.

Audience dependence creates challenges. For example, what is the relation between an expert explanation, which provides the best available understanding of a certain phenomenon, and an explanation of the same phenomenon for laypeople? A tempting answer is that a layperson explanation is an explanation sketch, or an approximation, of a full explanation, where the sketch provides the gist of the full story (the idea of sketches of historical explanations goes back to Hempel 1942). The problem with this solution is that if it is not communicated how a sketch deviates from the full story, the layperson explanation can be misleading, because it leaves out certain aspects of the full story. If, however, it is communicated how a sketch deviates from the full story, it can lead to the kind of information overload the approximation was designed to avoid. A different approach would require that a layperson is given an explanation sketch, or an approximative explanation, together with expert guidance on the difference between the sketch and the full story, depending on the goals of the layperson. This approach has the advantage that the gap between sketch and full story is bridged in a customized manner. A drawback is that it is much more costly than a "one size fits all" sketch.

Finally, radical context dependence means that the context dictates not only how the request for an explanation should be interpreted, but also what constitutes an admissible answer. A version of this problem was raised by Kitcher and Salmon (1987) as an objection against van Fraassen's pragmatic theory of explanation; the account given here draws on Woodward and Ross (2021). Even the kinds of facts that form the basis of an explanation can be chosen more or less freely. Radical context dependence does not seem adequate. Take the example of a request for information about close-by inputs that flip a prediction, (Q6) above. Not any kind of information is explanatorily relevant to answer this question, even if the notion of distance is not completely rigorous. Relevant information concerns the decision boundary of the predictor, plus, possibly, further information about the predictor F. This means that an adequate answer to (Q6) will have to draw on these facts, which means that what constitutes an adequate explanation is not completely context dependent. We will return to the discussion of the subjectivity and objectivity of explanations in Sec. 8.

6 Theoretical Explanations

So far, the examples of ML phenomena and explanation types were mostly situated at a concrete level of particular predictions. However, explanations

also occur at a more general and theoretical level. Higher generality means that the *explanandum* type is different. For example, we may ask about the behavior of an entire class of models, without reference to applications or single predictions. The tools to answer such questions are also different. If, say, our question is statistical and not tied to a particular dataset, the answer may be statistical or mathematical as well. Questions and answers of this type result in theoretical explanations.

A first example concerns the generalization properties of DNNs (see Zhang et al. 2021 on the concept of generalization, Kawaguchi et al. 2023 for a recent survey, and Buckner 2019; Räz 2022 for philosophical discussions). Many DNNs show a very good test set performance, that is, they generalize well. This phenomenon is general in that DNNs generalize well on may different datasets (e.g. image classification benchmarks), but also for different data modalities, such as images, text, and so on. This raises the question: Why do DNNs generalize well? This question is all the more relevant because *prima facie*, many DNNs are overparametrized, that is, they have many free parameters in comparison to the size of training sets, such that one would expect the models to overfit. This, however, is not the case. So far, there are several proposals and promising avenues, but no satisfactory answer to this why-question (Zhang et al., 2021). The generalization phenomenon is statistical in a double sense: The test set performance of single DNNs itself is a statistical property, and the generalization phenomenon concerns the frequency of this statistical property in applications of DNNs. This suggests that the eventual *explanans* will also be statistical in nature. In fact, many attempts to explain generalization behavior come from statistical learning theory (see the references given above). However, it is also possible that the *explanans* will include domain-specific knowledge, that is, a characterization of the kinds of features or data for which DNNs work well. There are different, interesting contrasts for the generalization phenomenon. One contrast is the question why the models perform well as opposed to overfitting the data. A second contrast, which may also be worth exploring, would be to better understand why DNNs do *not* perform well in some cases. The latter contrast may lead to a better understanding of the kind of features that contribute to good generalization.

A second example of a phenomenon requiring a theoretical explanation concerns the optimization properties of DNNs. The optimization (or learning) process used to train DNNs is usually a version of stochastic gradient descent (SGD). SGD is used to modify the weights of DNNs through backpropagation. There are many open questions regarding the optimization of DNNs. Usually, the optimization problem to be solved by DNNs is non-convex: if the error landscape is optimized locally using SGD, there is no guarantee that one ends up at a global minimum. Thus, the question arises: Why are DNNs optimized with SGD able to find global or close to global minima, as opposed to getting stuck in non-optimal local minima? This is a request for a general explanation. There has been interesting and relevant work on this problem, see e.g. Vidal et al. (2023), but it is still open, like the generalization puzzle. For example, the optimization properties of special kinds of simple networks (linear networks

with one hidden layer) have been explored, and it has been shown that for some kinds of data and architectures, optimization will not get stuck in local minima. This approach to explaining the optimization behavior of DNNs with "idealized" models can be interpreted as providing how-possibly explanations (Scholl and Räz, 2013; Verreault-Julien, 2019, 2023): it is established that the idealized model (a linear, shallow network) reproduces the behavior of the target system (deep, non-linear neural networks), for reasons that are mathematically well understood, but it is not clear to what extent the same explanations also apply to non-idealized models. More recent work has been geared towards de-idealizing these results, e.g., by extending the results to deep and non-linear networks.

A third example where an explanatory perspective may be useful is expressivity; see Gühring et al. (2023) for a recent review, on which the following draws. Expressivity explores the kinds of functions that DNNs can express in principle. Given a certain class of functions, such as continuous functions on a certain domain, it is investigated whether a class of models can approximate functions in that class. For example, one early universal approximation theorem states that a continuous function on a compact real domain can be approximated arbitrarily well by a neural network with sigmoidal activations (see Gühring et al. (2023) for details and Nielsen (2015) for a heuristic argument why DNNs are universal function approximators). Results of this kind can, again, be interpreted as providing how-possibly explanations, because they explain why these models are in principle capable of approximating a certain class of functions. However, expressivity results do not provide how-actually explanations. First, the process by which models are matched with functions does not correspond to the actual optimization procedures used in practice. Thus, there is no guarantee that one can approximate all functions in a class with the optimization procedures used in practice. Second, the model architectures used to prove approximation results can be highly idealized. For example, early universal approximation theorems with shallow networks needed the number of parameters to grow exponentially in the size of the input to achieve good approximations. These results are nevertheless useful, because they may suggest a path from highly idealized results to less idealized ones, and progress has been made since the early days. Note that other theoretical works, e.g., no-free-lunch theorems, may also have explanatory importance (see Sterkenburg and Grünwald 2021 for discussion of NFL theorems).

7 Levels

The examples of general explanations provided in the previous section suggest that one can explain ML phenomena at different levels of generality. For example, we can ask for explanations of single outputs, but also for explanations of statistical properties of generalization. If we accept that explanations come at different levels, the question arises as to how the levels are related. Below are some examples of relations between explanatory levels, highlighting their importance (see Dazeley et al. 2021 for a recent proposal regarding levels of

explanations in AI)

A first example can be gleaned from the discussion of contrastive explanations (Sec. 4). If we use contrast, we can formulate more focused requests for explanation and exclude certain possible answers as irrelevant. This also means that an explanation request without contrast is more general than the same request with contrast. For example, contrastive explanations that only consider particular values of the predictor function are special cases of an explanation of the global behavior of a predictor function. The relation between these explanations is deductive: global information about all input-output pairs entails information about specific counterfactual questions. Thus, if we are able to provide a global explanation of what a predictor function is doing, then this explanation encompasses everything we may want to know about counterfactuals, which means that the latter become superfluous. However, global explanations may not always be available. If a predictor function is too complex, we may not be able to grasp it, and we have to resort to approximations, or particular counterfactual scenarios, which are more tractable. Also, it is very hard, if not impossible, to specify formal criteria of functional interpretability, which corresponds to a global understanding of predictor functions (Räz, 2024). In a nutshell, the problem is that such understanding is possible if the corresponding predictors are simple, but the notion of simplicity itself is heterogeneous and has no formal characterization.

A second example is the relation between safe application of DNNs in particular contexts and our understanding of the generalization properties of DNNs. An explanation of the generalization properties of DNNs would be very useful because such an explanation would presumably reveal to what extent the application of these models allows us to understand or explain empirical phenomena. An explanation of the generalization properties might provide information about the contexts in which DNNs rely on spurious features, i.e., have high predictive accuracy "for the wrong reasons", and contexts in which DNNs do not rely on spurious features. Only in the latter case should we consider empirical information revealed by DNNs to be genuinely explanatory. The fact that we do not (yet) have a consolidated theory of deep learning means that the different explanatory levels are not firmly integrated yet (see Räz and Beisbart 2022, Sec. 5.).

8 Objectivity and Idealization

Are explanations of ML phenomena objective? The answer hinges on what we mean by "objective". It is helpful to tackle a different question first: In what sense could explanations be subjective? One possible source of subjectivity is the audience dependence of explanations (Sec. 5). Surely, an explanation is only successful to the extent that people are able to grasp it. If so, this commits us to a view of explanations according to which at least a certain degree of grasping (or understanding) is necessary for a successful explanation; more on this below. However, if this is taken to the point where grasping depends on the individuals

involved, then explanations become truly subjective. Compare this with scientific explanations: whether or not a (scientific or mathematical) explanation can be grasped is not a matter of individuals, but a matter of consensus in the scientific community. This does not make explanations completely objective: If what constitutes an explanation is determined by a community, the community can disagree about what can be grasped, and the community can be wrong. However, it also shows that explanations are not entirely subjective or up to individuals either. The problem of how to bridge the gap between explanations for experts and for laypeople was briefly discussed in Sec. 5.

The problem of graspability leads to a general feature of explanations. On the one hand, it should be granted that explanations do, in fact, serve a psychological purpose: they should be graspable in principle, after adequate scientific training, or with guidance. On the other hand, explanations do not reduce to what has been called a sense of understanding (Trout, 2002). They are grounded in facts, at least to some extent. Thus, explanations have two main ingredients, their psychological role, and factivity; see Wilkenfeld (2017) for a discussion of these two components in the debate on understanding, and Räz (2024) for an articulation of this idea in the context of understanding ML. If an explanation does not have one of these ingredients to a sufficient degree, it loses its explanatory status. To a large extent, the challenge of articulating good explanations boils down to the fact that these two ingredients are necessary, but also in tension.

To see how the tension arises, take the example of explaining single outputs of ML models (Sec. 4). In principle, we can simply cite the entire history of a model leading up to the output; this includes the entire model, training process, training data, and how each of these components were generated. This, however, is not an explanation because the entire history of an output is not graspable, it it is too large, which leads to cognitive overload. Thus, certain facts have to be left out or averaged over. Also, depending on the exact explanatory question, not all facts are relevant. However, even only relevant facts may lead to cognitive overload. Thus, usually, idealizations have to be introduced. Idealizations are deliberate distortions of facts, which allow us to grasp explanations. But every idealization compromises factivity. Thus, idealization is the second threat to objectivity.

Some critics have rejected approximative explanations as promoted by XAI for the very reason that they are necessarily wrong. In particular, Cynthia Rudin (2019) has argued that explainability methods are inadequate because they are non-factive. As a remedy, she has recommended to use interpretable models, such as rules lists, which are inherently and globally interpretable, while not compromising too much on accuracy. If it is really possible to build globally interpretable models that show predictive performance similar to black-box models, they should be preferred. The worry with this approach is that a loss of predictive performance will be incurred. Such a loss can be interpreted as a loss of a different kind of factivity, viz. the ability of the model to capture an empirical phenomenon. Note that recent results suggest that we can expect similar performance of black-box models and interpretable models at least for

tabular data (Chang et al., 2021). If we accept these results, then we should prefer interpretable models.

A second kind of example of explanations relying on idealizations are theoretical explanations (Sec. 6). The expressivity results of DNNs are mathematical results about the possibility of DNNs to represent certain function classes. At face value, such results are not idealized, they are mathematical statements, and contain no falsehoods. Considerations of idealization come in when we apply these results to DNNs with architectures as used in practice. Very often, such DNNs will not share some of the properties of the models that are investigated theoretically. This raises the question as to how the theoretical results bear on cases in which the assumptions of a theorem are violated. Above, it was suggested that we can interpret such results as how-possibly explanations. However, it is not clear whether these can be turned into how-actually explanations, e.g., by generalizing the mathematical results to more realistic architectures. At least some recent efforts (cf. Gühring et al. 2023) can be interpreted as working towards such de-idealized results, including work on the role of depth of DNNs in expressivity.

All in all, the use of idealizations is standard practice in science, and need not be necessarily problematic. A minimal requirement for the use of idealizations is that they should be labeled as such. One danger of idealized explanations is that even if such labels are provided in the original proposal of the idealized explanation, such labels are often left out in subsequent work, such that idealized explanations are interpreted as how-actually explanations.

9 Pluralism

Interpretability is pluralistic if there is not one kind of explanation of ML phenomena, but many. Pluralism with respect to interpretability has been advocated in the computer science literature by Lipton (2018), and also in philosophy by Krishnan (2020). Sometimes, the advocacy has been accompanied by calls for a formal definition (Doshi-Velez and Kim, 2017), sometimes with scepticism as to the possibility of a unified notion of interpretability (Räz, 2024). In the above discussion, explanation types, contrastive explanations, audience dependence, and different levels of explanation all point towards a certain degree of heterogeneity. Independently of the situation in computer science, scientific explanations presumably are also heterogeneous. There are many different proposals of how to define what a scientific explanation is, both in general and in particular scientific disciplines (see Woodward and Ross 2021 for scientific explanations and Mancosu 2018 for mathematical explanations). Some proposals may aspire to be generally applicable, but there is no generally accepted theory of what a scientific or mathematical explanation is, be it in the form of necessary and sufficient conditions, or in the form of more relaxed "theory of explanation". Some very general properties appear to be shared by most explanations, such as those used as a working definition above (Sec. 2.1), but these are too unspecific to merit being called a theory.

What does the state of the philosophical discussion mean for the discussion about explanations of ML phenomena? A critical stance towards the relevance of scientific and mathematical explanation for XAI and interpretability is defended by Páez (2019) and Krishnan (2020); see also chapter 9 in this volume. Even if scientific explanations are heterogeneous, this does not imply that explanations of ML phenomena are heterogeneous as well, simply because ML does not coincide with science (or math). However, the heterogeneity of scientific and mathematical explanations does have methodological consequences for the use of philosophical theories of explanation in application to ML. For example, if we can reconstruct a candidate explanation from ML in terms of a theory of explanation from philosophy of science, this does not imply that this candidate is therefore a good explanation, simply because there is no consensus in philosophy of science about what constitutes a good explanation. Such a reconstruction may be fruitful in some cases, but not in others. For example, showing that some mode of reasoning conforms with the DN theory of explanation is not useful per se, because it is known that some candidate explanations that have the form of a DN explanation are actually not very good explanations – put bluntly, the DN theory does not provide a useful picture of scientific explanations. For criticism of applying the DN theory in the context of ML, as proposed by Erasmus et al. (2020) see chapter 9 in this volume.

Now, even if there is no unified notion of explaining ML phenomena, this does not need to lead to a trivialization of the notion. One could hope that that for each kind of explanation, we can give a reasonably clear account of why it is an explanation. Even if there are several kinds of explanations, it is still possible that we end up with a limited set of reasonable kinds of requests for explanation, and with a limited set of acceptable answers to each request (see Räz (2024) for elaboration, and Durán (2021) for the need of a unified picture of scientific XAI). Of course, right now, such sets have not been identified. If this could be done, it would amount to what could be called a heterogeneous theory of explanations. Such a theory, because it is heterogeneous, would presumably involve distinctions between explanations along some of the dimensions mentioned above: types, context, contrast, levels, and so on.

Finally, note that ML also has the potential to play a unifying role in providing explanations of empirical phenomena, by generating explanations *with* ML. If progress with our theoretical understanding of ML (Sec. 6) were made, e.g., with an explanation of the generalization properties of ML, this explanation may tell us something about the features shared by domains in which ML is successfully applied. This, in turn, might lead to a non-trivial, theoretically-grounded account of the circumstances in which inductive reasoning, including scientific explanation, is successful, and also of the circumstances in which it is not.

10 Conclusion

To conclude, I would like to stress the advantages of viewing the problem of interpretability from the perspective of scientific and mathematical explanations. We can view these as the best explanations that science (including computer science, mathematics, etc.) has to offer. For some ML phenomena, there is no consensus as to what the best explanation would look like, and for other phenomena, there are no explanations altogether. In this situation, it can be misleading to provide explanations based on simplifications and visualizations that are merely graspable, but do not pay sufficient heed to factivity. It can be misleading because the simplified explanations may paint an inadequate picture of the best available explanation of a phenomenon, but also because the best available explanation may still be very imcomplete. Instead of communicating in terms of such simplified explanations, it is preferable to try to convey a picture of the complexity of the best available explanation, and also of the limits of our current knowledge.

Biography: Tim Räz is a postdoc at the Institute of Philosophy, University of Bern, Switzerland. A philosopher (PhD 2013, University of Lausanne) and mathematician (MSc. 2019, University of Bern) by training, he works on conceptual issues in machine learning and artificial intelligence, including interpretability and fairness.

Acknowledgements: I thank Claus Beisbart, Stefan Buijsman and Juan M. Durán for very valuable comments on earlier drafts of this chapter.

Funding: This work is funded by the Swiss National Science Foundation through grant number 197504.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6.

Beisbart, C. and Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, 17(6):e12830.

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

Bromberger, S. (1966). Why questions. In Colodny, R., editor, *Mind and Cosmos: Essays in Contemporary Science and Philosophy*. University of Pittsburgh Press, Pittsburgh.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, e12625.

Buijsman, S. (2022). Defining explanation and explanatory depth in xai. *Minds and Machines*, 32(3):563–584.

Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., and Caruana, R. (2021). How interpretable and trustworthy are gams? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 95–105.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4):568–89.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., and Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv:1702.08608v2.

Durán, J. M. (2021). Dissecting scientific explanation in ai (sxai): A case for medicine and healthcare. *Artificial Intelligence*, 297:103498.

Erasmus, A., Brunet, T. D., and Fisher, E. (2020). What is interpretability? *Philosophy & Technology*.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gühring, I., Raslan, M., and Kutyniok, G. (2023). *Theory of Deep Learning*, chapter Expressivity of deep neural networks, pages 149–99. Cambridge University Press.

Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition.

Hempel, C. G. (1942). The function of general laws in history. *The journal of philosophy*, 39(2):35–48.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15:167–79.

Kawaguchi, K., Bengio, Y., and Kaelbling, L. (2023). *Theory of Deep Learning*, chapter Generalization in Deep Learning. Cambridge University Press.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Kitcher, P. and Salmon, W. C. (1987). Van fraassen on explanation. *Journal of Philosophy*, 84:315–30.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33:487–502.

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.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521:436–44.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3):31–57. arXiv:1606.03490.

Mancosu, P. (2018). Explanation in mathematics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

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(3):441–459.

Räz, T. (2022). Understanding deep learning with statistical relevance. *Philosophy of Science*, 89(1):20–41.

Räz, T. (2023). Methods for identifying emergent concepts in deep neural networks. *Patterns*, 4(6). forthcoming.

Räz, T. (2024). Ml interpretability: Simple isn't easy. *Studies in History and Philosophy of Science*, 103:159–67. arXiv:2211.13617.

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

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–15.

Scholl, R. and Räz, T. (2013). Modeling causal structures. *European Journal for Philosophy of Science*, 3(1):115–32.

Shwartz-Ziv, R. and Tishby, N. (2017). Opening the black box of deep neural networks via information. arXiv:1703.00810.

Skow, B. (2016). Scientific explanation. In Humphreys, P., editor, *The Oxford Handbook of the Philosophy of Science*. Oxford University Press.

Sterkenburg, T. F. and Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese*, 199(3):9979–10015.

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

van Fraassen, B. (1980). *The Scientific Image*, chapter The Pragmatics of Explanation. Clarendon Press, Oxford.

Verreault-Julien, P. (2019). How could models possibly provide how-possibly explanations? *Studies in History and Philosophy of Science Part A*, 73:22–33.

Verreault-Julien, P. (2023). Three strategies for salvaging epistemic value in deep neural network modeling.

Vidal, R., Zhu, Z., and Haeffele, B. D. (2023). *Theory of Deep Learning*, chapter Optimization Landscape of Neural Networks, pages 200–28. Cambridge University Press.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31(2):841–87.

Wilkenfeld, D. A. (2017). Muddy understanding. *Synthese*, 194(4):1273–93.

Woodward, J. and Ross, L. (2021). Scientific explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34:265–88.

Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science*, 89(1):1–19.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.