Studies in History and Philosophy of Science 103 (2024) 159–167

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

ML interpretability: Simple isn't easy [☆]

Tim Räz

University of Bern, Institute of Philosophy, Länggassstrasse 49a, 3012 Bern, Switzerland

ABSTRACT

The interpretability of ML models is important, but it is not clear what it amounts to. So far, most philosophers have discussed the lack of interpretability of black-box models such as neural networks, and methods such as explainable AI that aim to make these models more transparent. The goal of this paper is to clarify the nature of interpretability by focussing on the other end of the "interpretability spectrum". The reasons why some models, linear models and decision trees, are highly interpretable will be examined, and also how more general models, MARS and GAM, retain some degree of interpretability. It is found that while there is heterogeneity in how we gain interpretability, what interpretability is in particular cases can be explicated in a clear manner.

1. Introduction

Machine learning (ML) models, and deep neural networks (DNNs) in particular, are very successful at solving problems both within and outside of science. However, many of these successful models are black boxes. The interpretability of ML models – understanding or gaining insight into how they work – is an important area of research in computer science.[1] Philosophers have started to pay more attention to interpretability; see Beisbart and Räz (2022) for a survey. Some philosophers have discussed more theoretically-oriented approaches (Buckner, 2019, Räz, 2022, Sterkenburg & Grünwald, 2021); there have been proposals for frameworks of explainable AI (Zednik, 2021); and it has been discussed whether understanding ML models is relevant to their usefulness in application (Sullivan, 2022, Räz & Beisbart, 2022).

The goal of the present paper is to explore the prospects of explicating the concept of interpretability in precise and unified terms. If we want to increase the interpretability of models such as DNNs, we have to get clear on what interpretability is. Computer scientists like Lipton (2018) have noted that interpretability is not a "monolithic concept"; Doshi-Velez and Kim (2017) call for a more "rigorous" notion of interpretability. Philosophers like Krishnan (2020) concur that interpretability lacks a clear meaning and question whether interpretability is an important problem in its own right.

So far, philosophers have approached interpretability with a focus on black-box models like DNNs. The present paper takes a different approach. Instead of focusing on the black-box end of the spectrum, where models lack interpretability, the focus here is on the other, interpretable end.[2] In computer science, interpretability has a tradition that predates the recent ascent of DNNs. The present paper follows the traces of this tradition in order to get a clearer picture of interpretability. Four interpretable ML models will be examined, with a focus on the properties that make these models interpretable, and how a higher degree of generality affects interpretability in two cases.[3] The pessimistic upshot of the paper is that even if we focus on a minimal notion of interpretability, it is still heterogeneous. This heterogeneity is traced back to four dimensions along which interpretability varies. The optimistic upshot is that there are common themes to the interpretability of different models, and that it is possible to explicate, in a reasonably clear manner, what it means to have a certain degree of interpretability for particular classes of models.

Section 2 introduces important aspects of ML models, and spells out how the debate on (scientific) understanding will be put to work to clarify interpretability. In section 3, linear models and (binary) decision trees, two interpretable regression models, are discussed. The properties that make these models interpretable are compared to identify core properties of interpretability. Section 4 turns to the question of how in-

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

E-mail address: tim.raez@posteo.de.

[1] Efforts to understand ML models in computer science run under different names. One kind of effort is towards what is called a theory of deep learning (Berner et al., 2023, Bahri et al., 2020). Another kind of effort, explainable AI (xAI, see, e.g., Adadi and Berrada 2018) aims to provide ML practitioners with tools to understand predictions made by the ML models they deploy.

[2] This approach is inspired by the distinction, stressed by Rudin (2019), between designing "inherently interpretable" models as opposed to applying xAI methods to opaque models.

[3] There are many useful discussions of interpretability in Hastie et al. (2009); the four cases discussed below can be found there. See Rudin (2019, 2021) for more on interpretable models, and Molnar (2020) for a book-length introduction, including more examples of interpretable models.

https://doi.org/10.1016/j.shpsa.2023.12.007
Received 21 February 2023; Received in revised form 14 November 2023; Accepted 16 December 2023
Available online 3 January 2024

terpretability scales with the generality of models. MARS and GAMs, two regression models that retain a certain degree of interpretability are examined, with a comparison of the properties that make the two models interpretable. Section 5 discusses general lessons about interpretability to be learned from the four cases. Section 6 concludes.

2. Preliminaries

2.1. Aspects of ML

The focus of the paper is on supervised learning (Hastie et al., 2009); other paradigms of machine learning like unsupervised learning and reinforcement learning are neglected. In supervised learning, we start with a dataset $\mathcal{D} = \{(x_i, y_i), i = 1...n\}$, sampled from an unknown distribution $P(X, Y)$, with instances x_i of X (inputs), labeled by instances y_i of Y (outputs). The variables X and Y can be continuous or discrete. The focus here is on regression problems, i.e., both X and Y are assumed to be continuous. The goal of supervised learning is to find a function $f(X) = \hat{Y}$ that is able to make predictions $f(x_i) = \hat{y}_i$ which are close to the "true values" y_i according to some loss function. Crucially, the accuracy of f is tested on samples from $P(X, Y)$ that were not used to construct f. If f performs well on such samples, it generalizes well. The goal of finding f is achieved by specifying a model M that computes f, and an optimization procedure, or learning algorithm, that adapts the parameters of the model M in a learning process, such that the model approximates the relation between the x_i and the y_i in $\mathcal{D}$. In what follows, f will be assumed to be a mathematical function, and variables to be ranging over sets; the probabilistic perspective, in which X, Y are random variables and f a distribution, will be neglected.

The concept of interpretability explored in this paper – the degree to which we understand an ML model – focuses on some aspects of supervised learning, while bracketing others. First, we can try to understand an ML model itself, or we can try to understand something about the world with the model. In the case of understanding *with* a model, the model plays an instrumental role in understanding something about the system from which the data $\mathcal{D}$ is sampled. In the case of understanding *of* a model, we are trying to understand the model itself. Here the focus is on understanding *of* a model; the question whether an ML models faithfully captures aspects of the world is bracketed.[4] Second, we can try to understand how a model arises through training, or we can try to understand a fixed trained model; see Räz (2022) for a discussion of the former. Here the focus is on the latter, i.e., understanding a fixed, trained model, or a family of such models. Finally, we can try to understand the inner workings of a model, or we can try to understand the function f computed by the model. Here the focus is on understanding the predictor function f. Understanding the inner workings of a model, its algorithmic properties etc. is important, and sometimes, there is no absolute distinction between model and predictor. Here, the inner workings of a model will be considered insofar as this serves the purpose of understanding the function f.

In short, the kind of interpretability to be investigated here, dubbed *functional interpretability*, is concerned with understanding the predictor function f computed by an ML model. This notion has been discussed in the philosophical literature on ML, in particular in the normative framework for xAI proposed by Zednik (2021). In Zednik's terminology, the goal is to understand *what* the predictor is doing, as opposed to understanding *how* (in terms of process) or *why* (in terms of a representational relation between model and world). In terms of the kinds of transparency distinguished by Creel (2020), the notion to be discussed here is a variety of functional transparency. Functional interpretability concerns the behavior of a predictor function as a whole, and thus requires a *global* kind of understanding, as opposed to local notions, which focus on understanding (or explaining) single predictions; the latter case is the usual setting of xAI methods.[5]

Why is functional interpretability important? First, because the main goal of supervised learning is to make predictions; therefore, it is crucial to understand how a model is behaving for various inputs. While understanding what an ML model predicts will not tell us everything about that model – it is not sufficient for full-blown understanding – it is arguably necessary: we do not understand an ML model unless we know what it does (to some extent). This is so because, e.g., understanding *why* a model behaves in a certain way presupposes some understanding of what the model does. Understanding why concerns the relation between the predictive behavior of a model and the model's target system in the world. However, we can only understand this relation if we have some prior knowledge or understanding of the predictive behavior of the model, i.e., some functional interpretability. Second, and relatedly, understanding what a model is doing is probably the aspect of ML models that affects most stakeholders – it is not only of interest to model builders, or model users, but also to decision subjects and policymakers. For example, in order for decision subjects to be able to challenge predictions that impact them, they have to know what these predictions are in the first place. Therefore, a clear concept of understanding this aspect of ML models is of high practical relevance.

2.2. Interpretability and scientific understanding

In the last section, the object of interpretability, the function f, was specified. This section discusses the notion interpretability itself. The idea is to explicate interpretability through understanding. It is assumed that a model with high interpretability is a model for which the degree of understanding is high. Thus, it should be spelled out what the degree of understanding is. To do so, we draw on discussions of understanding from philosophy. Understanding is arguably one of the central goals of science (de Regt & Dieks, 2005) and has been discussed in philosophy of science and epistemology. For a long time, understanding was seen as a mere psychological by-product of explanation (Woodward & Ross, 2021); only more recently has it been recognized as an achievement in its own right.[6]

For present purposes, understanding is taken to have several distinct but related dimensions (see also Wilkenfeld 2017). First, the agents or subjects who want to understand a model (or its representation) need to be able to *grasp* this model, it needs to be intelligible to them. The grasping dimension of understanding is not purely subjective, it does not reduce to a sense of understanding (Trout, 2002). The grasping of a model should be such that it can be taught, acquired, and verified an intersubjective manner. We do not want to simply replace interpretability with grasping; this would only move the challenge of spelling out interpretability to grasping. In order to proceed, concrete, operational criteria for grasping are needed. There have been different proposals in the literature for such criteria. One proposal is that the representation of a model can be grasped to the extent that it is possible for an agent to reason about the representation, manipulate the representation, and/or use it to make counterfactual inferences (Kuorikoski & Ylikoski, 2015). A second proposal is that a representation can be grasped (is intelligible) to the extent that an agent can anticipate qualitative consequences of the representation, without calculations or quantitative inferences (de Regt & Dieks, 2005), for example through visualization (de Regt, 2014). Below, these ideas will be

[4] In the philosophical literature, it has been discussed to what extent understanding *of* a model is relevant to understanding *with* a model, which seems the primary concern. Sullivan (2022) argues that our current understanding of ML models is sufficient to use them to understand the systems modeled; this has been challenged by Räz and Beisbart (2022).

[5] Functional interpretability is thus a more general notion that the explanation of particular predictions, as analyzed by, e.g., Watson and Floridi (2021), Zerilli (2022).

[6] See Baumberger et al. (2017) for a survey and Beisbart and Räz (2022) for a discussion of the relation between understanding and intelligibility/grasping.

put to use in the analysis of the models. We will see that manipulating a model and visualizing it are useful ways of grasping that model, often in combination. Further, important ways of grasping will be added to this list during the analysis.

A second important dimension of understanding is *accuracy*, the extent to which a representation allows us to grasp something about the system that is represented. In the present paper, the question of accuracy (degree of correspondence between model predictions and the world) is bracketed, because the focus is on understanding an aspect *of* an ML model itself, not understanding *with* a model, i.e., whether the model provides an adequate representation of something in the world. A third issue is whether understanding is taken to be categorical – we either have understanding, or we do not – or *graded*, that is, understanding is taken to come in degrees; see Baumberger (2019), Jebeile et al. (2021). Here a graded notion of understanding will be used. It will be argued that a graded notion should be preferred over a categorial notion in the analysis of the cases and in the discussion.

In sum, the object we want to understand is the function $f : X \rightarrow Y$ computed by a trained ML model M. We understand the function to the extent we grasp it, e.g., by observing how manipulating the input changes the output, or by examining how the function behaves qualitatively through visualizations. This proposal, which is still general and vague, will be refined in the discussion below.

3. Interpretable models

In this section, *linear models* and *decision trees* are examined; these two models are considered to have a high degree of interpretability by ML researchers. Properties that contribute to the intelligibility of these models are discussed. It is argued that while the two models share some of these properties, their formal properties and the way in which we grasp these models are so different as to yield two different paradigms of interpretability.

3.1. Linear models

Linear models are an important class of interpretable models. In “Elements of Statistical Learning” (ESL), a standard textbook, we find the following (typical) description: “[Linear models] are simple and often provide an adequate and interpretable description of how the inputs affect the output” (Hastie et al., 2009, p. 43). A linear model of n (continuous) input variables $X = (X_1, ..., X_n)$ and one continuous output $Y = f(X)$ is of the form

$$f(X) = \beta + w_1 X_1 + w_2 X_2 + ... + w_n X_n, \tag{1}$$

where we have added the intercept β.[7] The model parameters $\beta, w_1, ..., w_n$ are estimated in the learning process. Fig. 1 provides an illustration.

Why are linear models considered to be interpretable? Here are some relevant properties.

1. The linear model has a simple geometrical meaning: it corresponds to a hyperplane over the input space (Fig. 1). Additionally, such hyperplanes can be easily visualized for one- and two-dimensional inputs. Note that while models with more variables cannot be fully visualized, it is implausible that linear models with more variables are fully non-interpretable for that reason.
2. Model parameters have an intuitive meaning: The w_i correspond to the strength with which the input X_i is weighted in the computation of the output. For linear models, the model parameters directly translate into how the predictor function f behaves.[8]
3. Linear models are additive, which means that there are no interactions between input variables. Once we know the contribution of individual features, the overall output is a sum of these contributions, i.e., additive. This is a weak form of decomposability.[9]
4. For linear models, the same change in an input dX leads to the same change in the output dY everywhere. Or, put differently, the local shape of the function is also its global shape. This property uniquely characterizes linear models.[10]

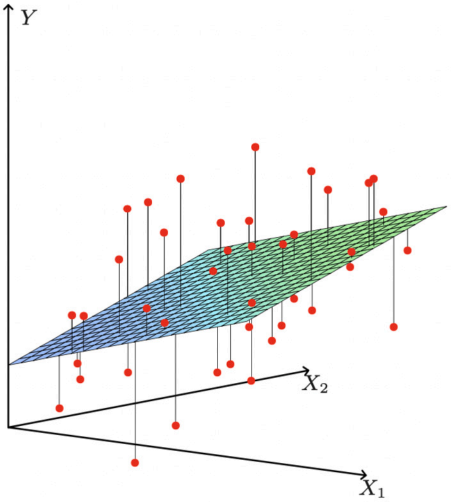

Fig. 1. Linear model with inputs X_1, X_2 and output Y; the (red) dots are the data points to be approximated by $Y = f(X_1, X_2)$. From Hastie et al. (2009), © 2009 Hastie, Tibshirani & Friedman. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Note that this list may be incomplete, and that the properties are not mutually exclusive. Consider how these properties fit into the discussion of grasping models from the previous section. The first property suggests that linear models are graspable through visualization; the fourth property may be recast in terms of grasping through local manipulation. The second and the third property suggest a way of grasping not mentioned above: we grasp linear models through the *form of the predictor function*. Of course, it is not surprising that linear models have these formal properties – this is how they are defined. The point is that grasping linearity encompasses the grasping of form: by using mathematical notation, we can *see* that the form of a function is linear. The suggestion here is that all of the above properties contribute to our understanding of linear models: we grasp through formal properties, as well as geometrical interpretation, visualization, and local behavior.[11]

It is not universally accepted that linear models are interpretable *tout court*. Lipton (2018) states several reasons why linear models are only interpretable with some qualifications. First, if the input space is

[7] Strictly speaking, the *function* f is affine rather than linear due to the β term, but this issue will be ignored here.

[8] Lipton (2018) calls this *decomposability*. This property distinguishes linear models from, say, DNNs, where the relation between model parameters and output is not straightforward.

[9] Additivity is a weaker property than linearity; not all additive models are linear; see the discussion of generalized additive models in section 4.2 below.

[10] This is true because linear models are the only models with constant derivatives everywhere. Note that we subsume constant functions under linear ones because we do not require the intercept β to be zero.

[11] From a computational point of view, it should be added that there are efficient and well-known procedures to fit a linear model to data. This makes sure we are able to find a linear model in the first place. Also, given any input, we can compute the corresponding output of the model in an efficient manner. Lipton (2018) calls this *simulatability*.

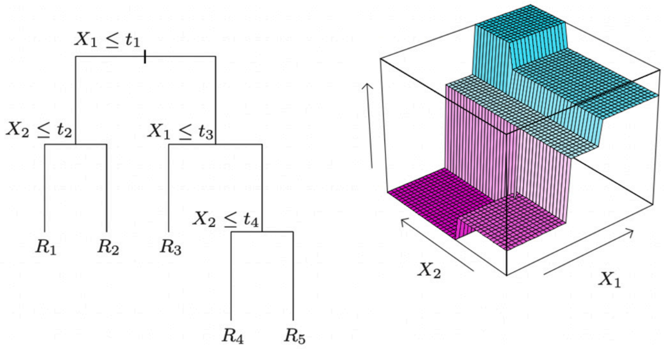

Fig. 2. CART: A binary regression tree in two variables (left) and the corresponding regression function (right). From Hastie et al. (2009), © 2009 Hastie, Tibshirani & Friedman.

high-dimensional, then carrying out computations, be it finding a model or computing particular outputs, becomes harder and harder. What is more, it may also be hard to gain an overall picture of how inputs affect outputs for high-dimensional models, and we may lose our overall grasp of the model's behavior, even if we understand how linear models work in general.

These qualifications are reasonable. However, they do not apply to linear models exclusively, they are a general feature of interpretability: The interpretability of models decreases as the dimension of the input space increases.[12] It seems natural to say that, generally speaking, the *degree* of interpretability decreases as the dimension of the input space increases. This is compatible with the possibility that we may lose our grasp of a certain kind of model above a certain threshold of the input dimension.

Lipton (2018) raises further points: linear models are fragile in that they sensitively depend on the selection of input variables, and also on pre-processing data, and, in order to get a degree of accuracy from linear models that is comparable to more complex models, linear models require pre-processed or pre-engineered features, which, in turn, may compromise interpretability. These arguments are valid in substance; the sensitive dependence on variable selection is a real problem. However, these are issues of understanding *with* the model, or how the model and the world are related, which are bracketed here. Importantly, even in cases where stability is not an issue, we still need a clear notion of understanding a given predictor function.

3.2. Decision trees (CART)

The second family of interpretable models are decision trees. Here we consider a simple kind of decision tree, so-called *classification and regression trees* (CART). Decision trees are taken to be interpretable by many computer scientists. Hastie et al. (2009) write: "Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. They are conceptually simple yet powerful" (p. 305).[13] Here is a short description of CART.[14] CART are recursive, binary decision trees that correspond to partitions of the input space into regions. The regions are obtained by recursively splitting the range of variables; see Fig. 2 for an illustration.

We start with training data $\mathcal{D}$, defined for variables $X = (X_1, \ldots, X_n)$ and Y. In the first step, we search through all possible partitions of the input space into two regions R_1, R_2. The two regions have the form $R_1 = X_i \leq t; R_2 = X_i > t$, that is, we split each variable X_i at values t of datapoints in $\mathcal{D}$. We choose the variable X_i and the split t such that, if we calculate the average value of the inputs in the two regions, and compare the result with the data, we get a minimal empirical loss. The recursive procedure is continued in the different regions until a certain tree size is reached; afterwards, the tree is pruned (internal nodes are collapsed) until a given tradeoff between model complexity (size of the tree) and predictive accuracy is satisfied. The resulting predictor function $f : X \to Y$ can be written as follows:

$$f(x) = \sum_i c_i \cdot I(x \in R_i). \tag{2}$$

Here, R_i, $i = 1 \ldots m$, are the regions of the partition, $c_i \in Y$ the prediction values in these regions (the average value of $\mathcal{D}$ in region R_i), and I the indicator function: 1 for $x \in R_i$, and 0 else.

What are the properties that make decision trees (CART) interpretable?

1. CARTs have a simple geometrical interpretation, they correspond to functions that are constant over (simple) regions of the input space; see Fig. 2 above. CARTs also allow for intuitive visualizations as trees, and the corresponding partitions can also be visualized, at least for low dimensions. Visualizability is very restrictive as a necessary requirement for interpretability, as in the linear case.
2. The regions R_i with constant predictions have a simple description in terms of the input variables. Hastie et al. (2009) consider CART to be particularly interpretable for this reason. They note that if we were to consider any (rectangular) partition of the input space, the resulting regions could be complicated to describe. This problem is solved by using recursive binary partitions.[15]
3. The prediction of CART is based on a sequence of binary decisions, which is easy to grasp. This means that the model represents a simple prediction process.[16] Note that this is a property of how the model processes the input, whereas the goal here is to characterize the interpretability of the resulting predictor function.

According to these properties, the interpretability of decision trees is closely tied to their representation. For one, it is important to grasp the partition associated with the decision tree. In the case of CART, the partition is grasped through the binary tree, which provides us with simple descriptions of the partition regions R_i: Every region is characterized through the sequence of splits at the internal nodes leading to R_i. This description is implicit in the form of the predictor function (2) via the regions R_i. For decision trees, visualization is also important, but it plays a different role than in linear models: the visual representation of the tree helps us grasp the structure of the partition associated with it. Also, a tree visualization (cf. left of Fig. 2) may help us grasp a function of more than two variables, where a visualization of the graph of a function may not be available. Thus, grasping a decision tree appears to work quite differently from grasping a linear model; a systematic comparison follows in the next section.

Decision trees, like linear models, are not interpretable without qualification. First, Lipton's argument about the dimension of the input space applies to decision trees as well. An important difference between linear models and decision trees is that we may be able to grasp a small decision tree even if the input space is large. A small decision tree is a tree with few splits, which means that the variables that are not split do

[12] To counteract a loss of interpretability in higher dimensions in the linear case, one can construct sparse models (models with dependence on few variables), e.g. using LASSO regularization, and use variable selection methods. One purpose of these techniques is to make high-dimensional models more interpretable, cf. Hastie et al. (2009, Ch. 3).

[13] See also Rudin (2021), Lipton (2018) for similar assessments.

[14] The presentation here follows Hastie et al. (2009, Sec. 9.2).

[15] "A key advantage of the recursive binary tree is its interpretability. The feature space partition is fully described by a single tree. With more than two inputs, partitions [...] are difficult to draw, but the binary tree representation works in the same way" (Hastie et al., 2009, p. 305).

[16] "[The tree representation] is also popular among medical scientists, perhaps because it mimics the way that a doctor thinks. The tree stratifies the population into strata of high and low outcome, on the basis of patient characteristics" (Hastie et al., 2009, p. 305).

not contribute to the prediction and can be ignored. Note that the same can be said about sparse linear models.

Second, decision trees can overfit the data if they become too large. If no stopping condition is used, and the tree is allowed to grow indefinitely, a tree can fit any (finite) dataset perfectly. This means that a tree can be made to agree with arbitrarily complex functions. However, such functions are not interpretable, and neither are trees that can be made to agree with such functions at any finite set of points. Thus, if the size of trees is not limited, they are not intrinsically interpretable. This feature of decision trees is not shared by linear models, which do not overfit the data to the same extent. Thus, the interpretability of decision trees has to be qualified: It applies to small trees only. This is, again, a point where a graded notion of understanding is important: With a categorical notion of interpretability, it would be necessary to say what "small" means, whereas on a graded notion, we can say that interpretability decreases as the tree grows. Third, decision trees can be sensitive with respect to small changes in the training data; see, e.g. Hastie et al. (2009, p. 312). As in the case of linear models, this problem is outside the scope of the concept of functional interpretability explored here.

3.3. *Linear models and trees: two paradigms of interpretability*

Now we examine whether there is a common explication of the interpretability of these two kinds of models. If interpretability is a monolithic concept, we should be able to identify common properties of these two highly interpretable models, and spell out what makes a model highly interpretable on this basis. If interpretability is not monolithic, but heterogeneous, we may still be able to characterize the main features that make these kinds of models interpretable separately, but without overlap of the main properties.

One way in which we might explicate interpretability is via common mathematical properties of predictor functions. This, however, does not seem to work: If we want to assign a high degree of interpretability to both linear models and decision trees, then highly interpretable predictor functions are not necessarily a) linear, b) differentiable or smooth, c) continuous, d) monotone, because small decision trees lack all of these properties in general.[17] Thus, there is no straightforward characterization of a high degree of interpretability through mathematical properties of predictor functions. Defining a monolithic concept of interpretability with these mathematical properties does not work. There is a very limited class of functions that belongs to both linear models and decision trees, the functions at the "intersection" of the two families: a decision tree with no splits, which corresponds to a linear model where all coefficients except for the intercept are 0. These globally constant functions are maximally interpretable. They compress the data into one number, which can be chosen to be the mean value or the median of the data.

Now let us examine the properties that make the two kinds of models interpretable. Both linear models and decision trees allow for visualization in low dimensions, and both models have a simple geometrical interpretation. However, many aspects of the interpretability of the two models are different. First, linear models are grasped through the *form of the predictor function*: the form of the linear function means that the (constant) contributions of individual variables to the output can be considered separately and contribute additively to the output. The predictor function of a decision tree, on the other hand, is grasped through the *form of the partition*, which is represented by the structure of the tree. If you want to grasp the predictor function of a tree, you need to grasp, first, how the tree partitions the entire input space (through the visualization of the tree), and, second, you need the values assigned to the regions of the partition to get outputs. A second, related dissimilarity is that linear models allow for local-to-global inference: once you know how a linear model behaves at one point, you know how it behaves everywhere. You can grasp a linear model through local manipulation. This is not true for decision trees. The behavior of the predictor function of a tree in one region does not tell you anything about its behavior elsewhere. Put differently, linear functions can be grasped bottom-up, while a decision tree is grasped top-down – you start with the entire partition and proceed to the values in the regions. A third dissimilarity is that decision trees are not intrinsically interpretable. To make them interpretable, a tradeoff between accuracy and simplicity has to be chosen. Linear models, on the other hand, intrinsically do not overfit the data.

These dissimilarities suggest that linear models and decision trees belong to two different paradigms of interpretability. On the linear paradigm, we grasp through the form of the predictor function, and interpretability is high because this form is simple. On the tree paradigm, interpretability hinges on the way in which we partition the input space, and a high degree of interpretability results from the fact that the form of the partition is simple (description/visualization as a binary tree). Thus, even on the minimal kind of functional interpretability considered here, interpretability is not a monolithic concept. The point is that we grasp the predictor functions in two different ways: via the representation of the function itself, and via the representation of the partition with a tree.

Let us return to the discussion of grasping in section 2.2 with these findings in mind. On the one hand, several of the criteria for grasping from the literature prove to be useful in the present context. We can grasp linear functions through local manipulation, and we can reason qualitatively on the basis of visualizations about both linear models and trees. On the other hand, some aspects of grasping a predictor function identified here do not feature prominently in the literature. Importantly, we grasp both linear functions and trees through the *form* in which they are represented, be it the predictor function or the partition associated with the predictor. From a mathematical point of view, this is not surprising: Understanding mathematical objects works as much through geometrical interpretation and visualization as through the formal, algebraic representation of the objects.[18]

4. Generalized interpretable models

In this section, MARS and GAMs are discussed. These are two more general models that retain a certain degree of interpretability. This will help us understand how the degree of interpretability changes with generality. The comparison between the models will show that a unified explication of interpretability is hard to come by, but that there is a reasonably clear way in which these models are interpretable if considered in isolation. The possibility of unifying the linear paradigm and the tree paradigm is explored, and it is argued that MARS provides a weak form of a unification of these two paradigms, with an explicit tradeoff between them.

4.1. *Multivariate adaptive regression splines (MARS)*

"Multivariate Adaptive Regression Splines" (MARS), a kind of regression model, combines (stepwise) linear regression and the CART regression tree model.[19] The model is built in two stages, a building

[17] Selbst and Barocas (2018) discuss to what extent so-called "inscrutable" functions lack properties such as linearity, continuity, and monotonicity. The argument made here suggests that a certain, low degree of "inscrutability" is compatible with functional interpretability. Note that it is assumed that the domain of trees is the feature space before the recursive partitioning step.

[18] The importance of combining principled, formal modes of reasoning with visualization in the context of data analysis is also stressed by Rosenstock (2021). Topological data analysis uses methods from algebraic topology to understand topological features of data clouds. Note that the primary application of this method appears to be cluster analysis, a form of unsupervised learning.

[19] The presentation here follows Hastie et al. (2009, Sec. 9.4).

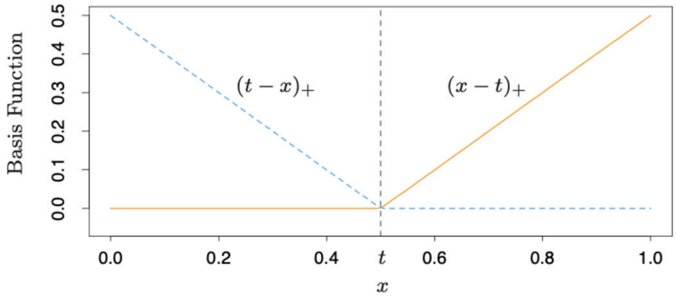

Fig. 3. Pair of basis functions (ReLUs) for MARS. From Hastie et al. (2009), © 2009 Hastie, Tibshirani & Friedman.

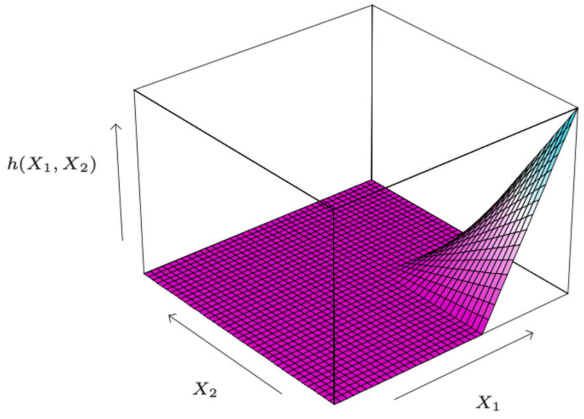

Fig. 4. A prediction function resulting from the MARS procedure; the function is $h(X_1, X_2) = (X_1 - x_{51})_+ \cdot (x_{72} - X_2)_+$; x_{51} and x_{72} are data points. From Hastie et al. (2009), © 2009 Hastie, Tibshirani & Friedman.

stage and a pruning stage, similar to regression trees. In the first stage, the model is built recursively from a set $C = \{(X_i - t)_+, (t - X_i)_+\}$ of pairs of piecewise linear functions in one variable, with knots t at the data points, see Fig. 3.

The starting point is the constant function. At each step of the building stage, two new terms are added to the model.[20] All possible products of pairs from C with terms already in the model are considered, and the two terms that lead to the largest decrease of the empirical error are added to the model. At the end of the building stage, the model consists of a sum of products of piecewise linear functions:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X), \tag{3}$$

where each of the functions h_m is one of the functions in C, or a product of such functions. The building stage terminates once the sum contains a preset number of terms. After the first stage, the model will usually overfit the data. In the second, pruning stage, the model is reduced by sequentially eliminating the term whose removal increases the empirical error the least. This yields a sequence of models of decreasing size. From this sequence of models, the final model is chosen through what is called generalized cross-validation, which selects the model that presents the best tradeoff between fit and simplicity – more on this below. An example of a prediction function resulting from the MARS procedure is shown in Fig. 4.

After this brief account of how MARS works, let us turn to its interpretability, which is one of the main objectives of MARS.[21] The property of MARS that contributes most to its interpretability is that its predictor function can be represented in a certain way.[22] The predictor function resulting from the construction process described above is of the form (3), which is not particularly telling. However, the terms in (3) can be rearranged to yield what Friedman (1991) calls an ANOVA decomposition:

$$f(X) = \beta_0 + \sum_{K_m=1} f_i(X_i) + \sum_{K_m=2} f_{ij}(X_i, X_j) + \sum_{K_m=3} f_{ijk}(X_i, X_j, X_k) + \ldots \tag{4}$$

β_0 is the intercept term. The index $K_m = 1$ runs over all functions h_m which contain one function from C; $K_m = 2$ runs over all functions that are a product of two functions from C, and so on. The functions $f_i(X_i)$ in the first sum consist of the piecewise linear functions from C in the variable X_i that enter into the model; the functions $f_{ij}(X_j, X_j)$ consist of all products of piecewise linear functions from C in the variables X_i, X_j that enter into the model, and so on.

How does the ANOVA decomposition (4) make MARS interpretable? First, the decomposition (4) reveals which variables enter additively into the model (and which ones do not), which pairs of variables enter quadratically into the model (and which ones do not), and so on. Second, at least the additive and the quadratic parts of (4) can be further investigated through visualization. If the degree of interaction is limited to two, this enhances the interpretability of the model significantly. If interaction terms are excluded, then MARS is additive, and thus even more interpretable, because plotting and interpreting one-dimensional functions is easier than interpreting two-dimensional ones.

Does MARS provide a "unification" of linear models and regression trees? The stepwise introduction of functions that are zero on part of the domain corresponds to a splitting or "breaking up" of the input space. At the same time, MARS generalizes stepwise linear regression, thus using aspects of the linear paradigm. In stepwise linear regression, the number of input variables in a linear model is controlled by only adding those input variables that contribute most to the accuracy of the model. MARS constructs the predictor function one variable at a time in the building stage.[23] These are aspects that MARS inherits from the two paradigms in the optimization phase.

The two paradigms appear most explicitly in the model selection during the second phase of construction. In this phase, the best model from the sequence of models f_λ is chosen, where λ corresponds to the number of terms in (3). For the choice of the best model, the criterion for generalized cross-validation (GCV) is used. GCV constitutes a tradeoff between predictive accuracy and simplicity. Both the use of many variables and the use of many splits make a model less simple according to the measure of complexity embedded in GCV. This means that the simplicity in MARS is itself not simple, but complex – it is literally a sum of two different "parameters of simplicity", one associated with the linear paradigm, one with trees. In this sense, MARS is an entire family of interpretable models, given by a parameter controlling for a tradeoff between two paradigms of simplicity. What is more, there is no theoretical justification for how to choose this tradeoff. The tradeoff is a pragmatic choice, which can be based on empirical accuracy. MARS

[20] MARS has several meta-parameters at the building stage that determine what kind of predictor function is constructed. First, an input variable can only feature once in a term. This means that there are no non-linear terms of one variable. Second, the degree of interaction terms, that is, the size of products can be chosen. If the upper limit is 2, then interactions are quadratic; if the limit is 1, then the model is additive.

[21] Other objectives are accuracy, smoothness, and computability; see the introduction of Friedman (1991). Note that Hastie et al. (2009, Table 10.1, p. 351) classifies MARS as an interpretable method.

[22] See the discussion in Friedman (1991, Sec. 3.5), on which the following discussion is based.

[23] Building sparse linear models is partially motivated by interpretability; see the discussion of stepwise linear regression in Hastie et al. (2009, Sec. 3.3.2).

thus inherits the two paradigms of interpretability, between which we have to choose.[24]

4.2. Generalized additive models (GAMs)

Generalized Additive Models (GAMs)[25] are a generalization of linear models. Linear models have the form (1), and thus assign constant weights to individual inputs. Additive models are more general in fitting smooth functions to individual inputs, while retaining additivity, that is, there are no interactions between inputs. An additive model has the following form[26]:

$$f(X) = \alpha + f_1(X_1) + \ldots + f_n(X_n). \tag{5}$$

The f_i may be non-linear, smooth functions, but it is also possible to choose some of the f_i to be linear and some to be non-linear, depending on what is known about the variables; GAMs are modular in this respect.

GAMs can be fit to data using the so-called backfitting algorithm. The idea of backfitting is to choose some initial values for the f_i, and then iteratively fit the f_i to the data using a procedure $S(f_i)$ until the fit stabilizes.[27] To elaborate, we first fit f_1 to the data in a smooth way, while keeping the $f_j, j \neq 1$ constant, then we fit f_2, using the new estimate of f_1 and the initial estimates for the remaining f_i, and so on; after one round, we start again with f_1, using the estimates for the $f_j, j \neq 1$ from the previous round. We continue estimating in this round-robin fashion until the differences between the f_i in two consecutive rounds are below a certain threshold.

Why are GAMs considered to be interpretable? Hastie and Tibshirani (1990, Sec. 4.3), who invented GAMs in the 1980s, note that GAMs are useful to analyze data. The form of the predictor (5) yields a generalization of the interpretability of linear models, which is based on additivity: If you change only one input variable, while holding all others fixed, the corresponding change in the output does not depend on the values of the other input variables. Hastie and Tibshirani write: “In practice this means that once the additive model is fitted to the data, we can plot the [functions f_i] separately to examine the roles of the predictors in modelling the response” (Hastie & Tibshirani, 1990, p. 88). Thus, the interpretability of GAMs is based on the fact that the predictor has the additive form (5) by design, and that, as a consequence, it is possible to examine the separate contributions of the f_i to the prediction by examining the plots (visualization) of the f_i.[28]

4.3. MARS and GAMs: compare and contrast

Which aspects of interpretability do MARS and GAMs have in common, and which are dissimilar? The two models are similar in how they achieve interpretability. In both cases, interpretability is a two stage process. In the first stage, we grasp through the form of the predictor function, a decomposition that shows what individual variables (or low-degree interactions) contribute to the prediction. In the second stage, the parts of this decomposition, the summands, are investigated separately, either through visualization (by plotting the individual functions of the decomposition), or by qualitatively investigating higher-degree interactions (in the case of MARS). By looking at the plots of individual functions, we can grasp qualitative aspects of the contributions of individual variables; of course, it is also possible to analyze the component functions in a more quantitative way.[29]

The two models are dissimilar in several respects. First, they constitute generalizations in different directions. GAMs are generalizations of linear functions in that they preserve additivity. They place no *a priori* restrictions on the form of the component functions, which can have almost arbitrary form (usually, they are nonparametric smooth functions). MARS, on the other hand, are geared towards discovering (low degree) interactions through the products of piecewise linear functions; this is a generalization in the spirit of trees. However, the component functions in the decomposition of MARS (basically low degree polynomials) are usually much simpler than the component functions in GAMs (nonparametric smooth functions). This facilitates an interpretation of the components of the ANOVA decomposition of MARS via the form of the component functions.

Second, MARS and GAMs differ in the way the predictor function is constructed, and, as a consequence, what the representation of the predictor function tells us. GAMs place the global constraint of additivity on the predictor function. There is no variable selection (at least in the automatic version), and all component functions are treated equally. MARS, on the other hand, is an adaptive procedure that automatically selects variables and interactions that are important for prediction. The ANOVA decomposition only contains variables and interactions that are important. The selective nature of the decomposition provides us with more information about the data – this is absent in GAMs. However, the ANOVA decomposition of MARS depends on the choice of the tradeoff in GCV, and this choice is pragmatic. The variable selection in MARS is at least partially determined by our preference for a more tree-like or a more smooth (linear) predictor function.

5. Discussion

Patterns of interpretability The four models exhibit some common patterns of interpretability. Most importantly, we grasp a predictor through a combination of two different means: First, the predictor function is grasped in virtue of the form of their representation (either the form of the predictor function itself or the form of the domain partition). Second, certain aspects, or parts, of this formal representation are grasped through visualization or other modes of qualitative reasoning. These two steps are used in sequence, not in parallel; this can be seen most clearly in the case of GAMs and MARS, but also in the case of trees. Thus, one important lesson for interpretability gained here is that at least some predictor functions are grasped by both looking at their visualization *and* inspecting their form, their algebraic representation, in combination.[30]

The importance of visualizations for interpretability is well known; they are an important part of explainability methods in xAI, cf. Molnar (2020). However, the fact that the formal representation of predictors (and the domain) is a key ingredient for interpretability is relatively unexplored. This may be related to the fact that little attention has been

[24] The tradeoff in ML between accuracy and simplicity is well-known both in computer science and philosophy; see, e.g., the discussion in Forster and Sober (1994) of this tradeoff with respect to model selection. The above discussion of GCV does not concern the tradeoff between accuracy and simplicity. GCV constitutes a three-way balance between accuracy and two ways of conceptualizing simplicity.

[25] The presentation here follows Hastie et al. (2009, Sec. 9.1).

[26] Here I gloss over some aspects of GAMs, such as the fact that GAMs may allow the response $Y = f(X)$ to be a smooth transformation of the sum on the RHS of (5), via a so-called link function.

[27] There are different possible choices for the procedure S, e.g. so-called cubic smoothing splines. What is important is that there are efficient procedures to find a good and smooth approximation of a non-linear function f_i in one variable.

[28] In the above description, GAMs are a fully automatic procedure. One problem with this approach is that the backfitting algorithm fits all variables, even those that do not contribute much to the output. It is possible to select and exclude variables by hand. GAMs are usually applied in this more interactive way in data analysis. However, there are also automatic procedures for finding sparse additive models (Hastie et al., 2009, Sec. 9.1.3).

[29] Note that visualization can be generalized to a certain extent and applied to functions with higher-degree interactions, e.g. through so-called *partial dependence plots*, cf. (Hastie et al., 2009, p. 369), and also Molnar (2020).

[30] There have been few discussions of the importance of the right kind of representation/notation in the context of mathematical explanation, cf. Colyvan (2012, Ch. 8), Räz (2018).

paid to formal dimensions along which interpretability varies, such as the size of the input space, the degree and complexity of interactions, and the nature of nonlinearities. The above discussion helps us to appreciate how these different dimensions interact and how we may have to trade them off against each other.

Four dimensions of interpretability From the above case studies, we can extract four dimensions that contribute to the degree of interpretability of ML models. First, generally speaking, the degree of interpretability decreases as the size of the input space grows. Second, and relatedly, the degree of interpretability increases as the size of the model decreases, e.g. by making a linear model sparse, or by controlling the size of a decision tree. Third, the degree to which we allow non-linearities in one variable affects interpretability. For example, using nonparametric smooth functions in GAMs results in less interpretable one-variable functions than piecewise-linear functions in the additive part in MARS. Fourth, interpretability decreases as we allow interactions of higher degree and complexity.

Recognizing these four dimensions of interpretability allows us to diagnose why it may be hard, or even impossible, to obtain a singular degree of interpretability. The reason is that while interpretability does vary along these four dimensions, there are complex tradeoffs between the dimensions that make it hard to assign a consistent, singular degree of interpretability. To give an example, it may be tempting to assign a degree of interpretability based on the degree of interaction terms, arguing that once we move beyond two-degree interactions, we lose the possibility to visualize, and thus the possibility to achieve interpretability. This suggests that we should define interpretability in terms of degrees of interactions. This, however, neglects the fact that even relatively small trees can exhibit higher degrees of interaction, while arguably still being interpretable. Trees are interpretable because their interactions are of a very simple kind – a sequence of binary decisions in the case of CART. Thus, interpretability does not simply decrease as the degree of interactions considered increases. A second example is the tradeoff implicit in the choice between MARS and GAMs, which is a tradeoff between using a global constraint on interactions while allowing for complex nonlinearities (GAMs), and allowing for some interactions, while restricting the form of nonlinearities (MARS).

The benign heterogeneity of interpretability The four case studies suggest that we should not expect a monolithic notion or degree of interpretability. Even the narrow notion of *functional interpretability* investigated here is heterogeneous. This strengthens a point made by Lipton (2018). Heterogeneity is witnessed by the linear paradigm and the tree paradigm. These do not only correspond to two kinds of predictor functions, but they also differ in the way we grasp them. The possibility of unifying the two paradigms in the form of MARS was discussed, and it was argued that MARS does retain aspects of both linear models and trees. However, the two paradigms do not vanish through generalization, but remain encoded in an explicit tradeoff in the GCM criterion. Note that the two paradigms are just one manifestation of heterogeneity. There are similar, subtler differences between the interpretability of MARS and GAMs. These points let us appreciate *why* functional interpretability is heterogeneous.

However, heterogeneity does not mean that we should not strive for interpretability as a goal in and of itself, *pace* Krishnan (2020). It also does not mean that interpretability is "ill-defined", *pace* Lipton (2018), or that anything goes. There are different means to obtain interpretability. But once these means are identified in particular cases, it is possible to get a clear sense of what we do understand about the corresponding models. To make an analogy, proving different mathematical propositions is heterogeneous as well, in that different methods may be necessary to establish different propositions. It is comparatively easy to check whether and why a given proof of a proposition is acceptable, and the criteria for checking this are also clear. Similarly, the ways in which we achieve interpretability are heterogeneous, they have to be adapted to particular kinds of models, and they are not yet available for many models. Still, the tools we do have to interpret GAMs, or MARS, are clear and understandable, and they do show some common patterns.

The prospects of a conceptual analysis of "interpretability" The idea that interpretability is a graded notion that varies along several dimensions is more general than the more traditional approach of identifying necessary and sufficient conditions for the concept of interpretability. If one is unwilling to give up on the project of identifying necessary and sufficient conditions, it is, in principle, possible to obtain a conceptual analysis on the basis of a graded notion of interpretability by introducing thresholds, i.e., by declaring that a model is interpretable if and only if it is interpretable at least to a fixed degree d. This presupposes that a singular degree of interpretability is available, and that a principled choice of threshold can be made. The challenge of coming up with such a singular degree of interpretability, discussed above, suggests that it will be even harder to find necessary and sufficient conditions for interpretability than to assign a degree of interpretability.

Ramifications for black-box models and xAI What are the ramifications of the above discussion for the interpretability of black-box models? If we try to locate black-box models such as DNNs along the four dimensions of interpretability, we have to assign them a low degree of interpretability, because they lie at the low-interpretability end of all four dimensions: They are usually applied in high-dimensional settings (e.g. image recognition), they have a lot of free parameters, and they are not only highly non-linear, but capture high-degree interactions. It is one of the main open puzzles about these models that they are predictively successful despite these properties. One of the messages of the present paper is that even if we focus on functional interpretability, there is not just one method of grasping a predictor function. Rather, we may need to employ different methods in combination for black-box models, as in the case of MARS and GAMs, where both formal properties of the representation of the predictor as well as visualizations play key roles in grasping.

Functional interpretability aims at global understanding (of an entire function) and is therefore different from local xAI methods (Adadi & Berrada, 2018), which rely on local and linear approximations in combination with visualizations. We can nevertheless glean some lessons from the above discussion for local explanations. For one, if we can gain a high degree of functional interpretability, our grasp of a predictor is global and should thus be preferred over local explanations of black-box predictions, other things being equal.[31] What is more, the discussion in the present paper suggests that obtaining visualizations of linear approximations is by no means the only way to gain interpretability. If the story presented here is correct, we can gain understanding of black-box models through a combination of understanding formal properties and visualizations. This is not to say that we will be able to use the methods from MARS, GAMs or trees on DNNs in a straightforward manner. Dealing with DNNs will require us to move beyond the linear paradigm and to discuss how we can grasp complex interactions in high-dimensional settings.

6. Conclusion

Interpretability was analyzed by examining four case studies of models with a certain degree of interpretability. There is a considerable heterogeneity with respect to the means by which we achieve *functional interpretability*. In particular, linear models and decision trees belong to

[31] Other things encompassing, in particular, the accuracy of a predictor. It is stressed, e.g., by Rudin (2019) that the use of interpretable models will not depreciate accuracy in many cases and should therefore be preferred over xAI methods.

two different paradigms of how interpretability is achieved. The same is true for the two more general, but still interpretable models, MARS and GAMs. However, interpretability is not ill-defined for these reasons. Rather, we can spell out clearly what interpretability amounts to in particular cases.

The above discussion is limited in several respects and should be extended. First, the case of classification problems, and of different kinds of inputs (discrete, mixed etc.) should be taken into consideration. Second, it would be desirable to extend the limited notion of functional interpretability explored here and relate it to existing analyses of interpretability, giving more weight to optimization, the representational role and the inner workings of models. Third, the investigation of interpretability should be extended to unsupervised learning, reinforcement learning, and other paradigms of ML.

Acknowledgements

I thank audiences in Bern, Tübingen and Munich and anonymous referees for useful feedback on previous versions of the present paper, and Martin Gasser for useful feedback and discussions. I also thank Trevor Hastie for permission to use Figs. 1–4.

References

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

Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., & Ganguli, S. (2020). Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, *11*(1), 501–528.

Baumberger, C. (2019). Explicating objectual understanding: Taking degrees seriously. *Journal for General Philosophy of Science*, *50*, 367–388.

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 epistemology and philosophy of science* (pp. 1–34). Routledge.

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

Berner, J., Grohs, P., Kutyniok, G., & Petersen, P. (2023). The modern mathematics of deep learning. 1. In *Theory of deep learning* (pp. 1–111). Cambridge University Press.

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

Colyvan, M. (2012). *An introduction to the philosophy of mathematics*. *Cambridge introductions to philosophy*. Cambridge: Cambridge University Press.

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

de Regt, H. W. (2014). Visualization as a tool for understanding. *Perspectives on Science*, *22*(3), 377–396.

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

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

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, *45*, 1–35.

Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. *Monographs on statistics and applied probability: Vol. 43*. Chapman & Hall/CRC.

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

Jebeile, J., Lam, V., & Räz, T. (2021). Understanding climate change with statistical downscaling and machine learning. *Synthese*, *199*, 1877–1897.

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

Kuorikoski, J., & Ylikoski, P. (2015). External representations and scientific understanding. *Synthese*, *192*, 3817–3837.

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

Molnar, C. (2020). Interpretable machine learning. Lulu.com.

Räz, T. (2018). Euler's Königsberg: The explanatory power of mathematics. *European Journal for Philosophy of Science*, *8*, 331–346.

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

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

Rosenstock, S. (2021). Learning from the shape of data. *Philosophy of Science*, *88*(5), 1033–1044.

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.

Rudin, C. (2021). Interpretable machine learning: Fundamental principles and 10 grand challenges. arXiv:2103.11251v2.

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, *87*, 1085.

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

Sullivan, E. (2022). Understanding from machine learning models. *British Journal for the Philosophy of Science*, *73*(1), 109–133.

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

Watson, D. S., & Floridi, L. (2021). The explanation game: A formal framework for interpretable machine learning. In *Ethics, governance, and policies in artificial intelligence* (pp. 185–219). Springer.

Wilkenfeld, D. A. (2017). MUDdy understanding. *Synthese*, *194*(4), 1273–1293.

Woodward, J., & Ross, L. (2021). Scientific explanation. In E. N. Zalta (Ed.), *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*, 265–288.

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