Minds and Machines (2022) 32:563–584
https://doi.org/10.1007/s11023-022-09607-9

ORIGINAL ARTICLE

Defining Explanation and Explanatory Depth in XAI

Stefan Buijsman[1]

Received: 11 June 2021 / Accepted: 14 June 2022 / Published online: 29 June 2022

Abstract

Explainable artificial intelligence (XAI) aims to help people understand black box algorithms, particularly of their outputs. But what are these explanations and when is one explanation better than another? The manipulationist definition of explanation from the philosophy of science offers good answers to these questions, holding that an explanation consists of a generalization that shows what happens in counterfactual cases. Furthermore, when it comes to explanatory depth this account holds that a generalization that has more abstract variables, is broader in scope and/or more accurate is better. By applying these definitions and contrasting them with alternative definitions in the XAI literature I hope to help clarify what a good explanation is for AI.

Keywords Explainable AI · Counterfactuals · Explainability · Manipulationism

1 Introduction

Artificial Intelligence (AI) based algorithms, especially using deep neural networks, are becoming ever more ubiquitous, introducing opaque decision systems into a wide range of applications. These algorithms can analyse large and complex data sets, setting state-of-the-art performance, but at the cost of interpretability. Deep neural networks in particular often have millions, or even billions, of parameters, making it near impossible to understand why a particular output was selected based on the (possibly also complex) input. Given the prevalence of such algorithms, however, it is important to be able to explain their outputs, also considering that the General Data Protection Regulation, accepted into EU law in May 2018, provides a right to explanation. The field of eXplainable AI (XAI) has taken up this challenge with a variety of tools (for reviews, see Adadi & Berrada, 2018; Das & Rad, 2020; Guidotti et al., 2018). Despite a plethora of approaches to solving the issue of how to explain the functioning of opaque decision systems (deep neural networks are one

✉ Stefan Buijsman
stefan.buijsman@gmail.com

[1] TU Delft, Jaffalaan 5, 2628 BX Delft, The Netherlands

example, but other machine-learning techniques such as random forests present the same issue), there is however no widely agreed upon definition of ‘explanation’, nor a way to compare the quality of different explanations.

Instead, one finds a range of definitions that often fail to provide specific guidance. For example: “We define explainability as the ability for the human user to understand the agent’s logic” (Rosenfeld & Richardson, 2019, p. 678). Similarly, Guidotti et al. (2018, p. 5) say that “Essentially, an explanation is an “interface” between humans and a decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans”. Somewhat more elaborately, Ciatto et al. (2020, p. 9) start with a definition of interpretation to define explanation: “we define the act of “interpreting” some object X as the activity performed by an agent A—either human or software—assigning a subjective meaning to X. Such meaning is what we call interpretation. [...] We define “explaining” as the activity of producing a more interpretable object X' out of a less interpretable one, namely X, performed by agent A”. These definitions, though certainly not wrong, do not yet tell us very much about what explanations are exactly. They do not state what an explanation should look like, nor what the essential features are of a statement that make it an explanation.

Opposed to these general definitions there are also more specific definitions. For example, Das and Rad (2020, p. 4) give the following definition: “An explanation is additional meta information, generated by an external algorithm or by the machine learning model itself, to describe the feature importance or relevance of an input instance towards a particular output classification”. This definition focusses on specific approaches that are currently followed, but doesn’t tell us why feature importance or relevance would explain the output of the algorithm, nor does it connect to a wider account of explanation. I aim to provide such an account here, by presenting the manipulationist definition of explanation from the scientific explanation literature in the philosophical of science. That is based on generalizations that answer what-if-things-had-been-different questions and is the most popular account of causal/scientific explanations outside of the XAI context. I motivate that it works well for XAI too by contrasting it with a few more definitions of explanation from the machine learning literature. As part of the comparison I also demonstrate how this independently motivated account of explanation might apply to tools used in XAI, such as saliency methods and rule extraction. Then, in Sect. 3, I discuss the question of explanatory depth, drawing again on the philosophical discussion of the topic in the context of scientific explanation. The goal there is to present a set of considerations that capture whether one explanation is better (deeper) than another. Primarily it will be a more precise specification of the idea that “powerful explanations should, just like any predictor, generalize as much as possible” (Guidotti et al., 2018, p. 36) drawing on the philosophical literature on explanatory depth. The goal of this paper thus is to make progress on what (Guidotti et al., 2018, p. 36) finds one of the most pressing issues in XAI:

> One of the most important open problems is that, until now, there is no agreement on what an explanation is. Indeed, some works provide as explanation a set of rules, others a decision tree, others a prototype (especially in the context

> of images). It is evident that the research activity in this field is not providing yet a sufficient level of importance in the study of a general and common formalism for defining an explanation, identifying which are the properties that an explanation should guarantee (Guidotti et al., 2018, p. 36).

2 Defining Explanation

2.1 The Manipulationist Definition

What is it that explains? A popular idea in the philosophical literature is that scientific explanations are all causal, i.e. that one explains a (scientific) fact by appealing to its causes (Salmon, 1984; Woodward, 2003). This accounts for a number of features of explanations, such as their asymmetry: one cannot both explain Y by appeal to X and X by appeal to Y. A standard example in philosophy to illustrate this point is of explaining the length of a flagpole and its shadow:

(1) Flagpole f has length l_1 because its shadow s has length l_2 and the sun strikes the flagpole with angle α

(2) Shadow s has length l_2 because the flagpole f has length l_1 and the sun strikes it with angle α

Explanation (2), but not (1), strikes us as a good explanation of the phenomenon in the real world, even though one can derive s from l_1 and α without any problems. Causal accounts explain this difference between (1) and (2) by pointing out that the sun striking the flagpole causes the shadow (and thus determine its length), whereas the shadow does not cause the flagpole to have a certain length. Hence, the real world phenomenon is best explained by (2), though (1) would be the better explanation for an *algorithm* that calculates the length of a flagpole based on the shadow and angle. This idea, that causes are what explain, is also found in part of the XAI literature. Miller (2019, p. 12) specifically offers a definition based on causation: “This paper adopts Lipton’s assertion that explanation is post-hoc interpretability. I use Biran and Cotton’s definition of interpretability of a model as: the degree to which an observer can understand the cause of a decision”. I agree with this definition to an extent, but I think that there is a more informative definition available by further specifying what is meant by ‘causation’.

The approach to this question that I take is similar that of Pearl and Mackenzie (2019) and Halpern and Pearl (2005a), as well as to his take on explanation Halpern and Pearl (2005b), though I will follow the account more commonly discussed in the philosophical literature, namely that of Woodward (2003). Watson and Floridi (forthcoming) also use this definition of explanation to give a formal framework of ‘explanation’ in the XAI context. My paper differs from theirs by looking not just at local explanations and by allowing explanations to be broader in scope than a single model (see Sect. 3 as to why this is an important improvement). Furthermore, their formal framework fails to make explicit how the manipulationist definition interacts with existing definitions and methods, and in what way it can act as a good goal for XAI methods to strive for. Therefore I think it is worthwhile to discuss the work of Woodward (2003) again as a good definition of 'explanation' in XAI, free from the strictures of a formal framework that casts XAI research in the form of an explanation game. To start on his account: he provides, first of all, a non-reductive definition of a cause based on what are called 'interventions', which informally are ways to change the value of a variable x without changing the value of the other variables that cause the purported effect y.[1] Using this notion, x causes y iff an intervention I on x changing the value from x_1 to x_2 produces a correlated change in the value of y from y_1 to y_2. That makes it inherently a counterfactual account: x causes y *if* an intervention on the value of x *would* change the value of y. There is no need to actually change these values for there to be causation, as all that matters is that the value of y depends on the value of x. It is, consequently, this dependence relation between the values of the cause and the effect that we are interested in when we look for an explanation, or so manipulationist theories claim. That marks a slight departure from Miller's definition, as it is not whether we understand the cause of a decision that matters (so this account holds), but whether we understand how the cause influences the effect that matters for an explanation. In other words, what we want to know about a black box algorithm is how the input determines the output.

These various strands are then combined into a definition of explanation, which basically maintains that an explanation is an answer to a range of what-if-things-had-been-different questions. Woodward writes that E explains M in the following case:

> Suppose that M is an explanandum consisting in the statement that some variable Y takes the particular value y. Then an explanans E for M will consist of
> (a) a generalization G relating changes in the value(s) of a variable X (where X may itself be a vector or n-tuple of variables X_i) and changes in Y, and
> (b) a statement (of initial or boundary conditions) that the variable X takes the particular value x.
> A necessary and sufficient condition for E to be (minimally) explanatory with respect to M is that (i) E and M be true or approximately so; (ii) according to G, Y takes the value y under an intervention in which X takes the value x; (iii) there is some intervention that changes the value of X from x to x' where $x \neq x'$, with G correctly describing the value y' that Y would assume under this intervention, where $y' \neq y$ (Woodward, 2003, p. 203).

[1] More formally, I is an intervention-variable on X with respect to Y if and only if: I1. I causes X. I2. I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I. I3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X. I4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X (Woodward, 2003, p. 98). An intervention $I = i_1$ is an instantiation of this intervention-variable that makes X take value(s) X_1.

The minimal case is very minimal here, as Woodward doesn't demand that the generalization G is error-free or covers more than one counterfactual case with a differing outcome. Those aspects of the definition are, however, a matter for the discussion on explanatory depth in Sect. 3. First I consider it helpful to contrast this definition with two more specific (formally specified) definitions from the XAI literature, and in doing so to apply the manipulationist account of (scientific) explanation to XAI. To do so, I propose an initial application of the definition to a (black box) algorithm b to be refined later on in the paper. An explanation of output y_1 of b, resulting from input X_1 is: a generalization G where $G(X_1) = b(X_1) \pm \delta$, with δ a chosen minimum accuracy of G. Furthermore, there is at least one set of inputs X_2 where $G(X_2) = b(X_2) \pm \delta$ and $b(X_1) \neq b(X_2)$. Importantly, the causes here refer not to the real-world phenomena that an algorithm might try to predict, but are purely about the relation between inputs of the algorithm and the outputs—thus they may reverse the 'natural' causal order provided they match what happens inside the algorithm. Explanations, on this picture, are thus rules that include counterfactual cases, and as such can answer what-if-things-had-been-different questions. That ties together different ideas already in the XAI literature, to which I turn now.

2.2 Counterfactuals Alone

A first useful contrast is with the definition of Wachter et al. (2018), which also relies on counterfactuals. They define an explanation of the output of an algorithm as having the following format: "Score p was returned because variables V had values $(v_1, v_2, ...)$ associated with them. If V instead had values $(v'_1, v'_2, ...)$, and all other variables had remained constant, score p' would have been returned. Wachter et al. (2018, p. 848) That definition is in line with the literature on algorithmic recourse (Karimi et al., 2021), which formalizes it further: "Given a fixed predictive model, commonly assumed to be a binary classifier, $h : X \rightarrow 0, 1$, with $X = X_1 \times \cdots \times X_D$,we can define the set of *contrastive explanations* for a (factual) input $\mathbf{x}^F \in X$ as $E := \{\mathbf{x}^{CF} \in \mathcal{P}(X) \mid h(\mathbf{x}^{CF}) \neq h(\mathbf{x}^F)\}$. Here, $\mathcal{P}(X) \subset X$ is a plausible subspace of X, according to the distribution of training data" (Karimi et al., 2021, p. 3) One can then select the closest element of E (based on a chosen distance function; Karimi et al. (2021) discuss various options) in answer to a why-question by a user.

There is a crucial contrast between these two definitions and the manipulationist definition that I think is a promising candidate for capturing what XAI is after, though they all appeal to counterfactuals. Whereas Woodwards definition is based on a generalization G, which describes the correlation between the values of the explanandum x and the explanans y (or $h(x)$ on the terminology of Karimi et al. (2021)), there is no such generalization included in the definitions of Wachter et al. (2018) and Karimi et al. (2021). So, is a definition with this generalization more fitting than their simpler definitions employing only counterfactual cases?

Consider first a few examples from outside of XAI, to see how both definitions fare in a less technical context. When asked why the window broke after someone threw a baseball at it with velocity v_1, we might be provided with the following two answers:

(3) If the baseball had hit the window with the lower velocity v_2 the window would not have broken.

(4) Glass breaks when struck by objects travelling at velocities higher than v_2. The baseball struck at velocity $v_1 > v_2$ and therefore broke the glass window. Had it travelled at v_2 the window would have remained intact.

In (4) the first sentence gives generalization G, the second shows that $G(v_1) = y$ (i.e. the window breaking) and the third sentence presents the counterfactual case. I think it's clear that (4) provides a better explanation than (3), but the question is whether (3) is also a good explanation. Technically speaking (3) would be preferable, as finding these counterfactuals is a well-defined problem that, though with its own difficulties, is easier than finding generalizations such as (4). However, I think that (3) only manages to explain something about the window breaking because it suggests a generalization that covers both cases: that there is a minimum velocity needed to break glass. In the absence of a clear generalization that we might infer from the counterfactual case it does not seem particularly explanatory to only specify a counterfactual. Consider the following explanation:

(5) If the Earth had mass m_2 then the window would not have broken.

Without further information on the physical reason for this fact (increase the mass of the Earth enough and the ball will drop to the ground before reaching the window due to increased gravitational effects, a suitable generalization here being Newton's laws) it provides very little illumination on why the window broke. Now, of course this isn't a close counterfactual by any measure, but such cases can easily occur with algorithms (though Kenny and Keane (2021) discuss generating plausible counterfactuals, and one could manually disallow using the Earth's mass in counterfactuals). My point is rather that a generalization, a sketch of which I just gave in parenthesis, is the crucial additional factor for the explanation to make (some) sense. The same holds within XAI. Here too some contrasts will seem explanatory because they strongly suggest a generalization. For example, when applying for a loan the contrast 'had your income been $x_2 > x_1$ then the loan would have been approved' strongly suggests a rule that states the relation between income and the maximum amount you can borrow. We infer that rule from the context and thus complement the counterfactual to reach a proper explanation. There is no guarantee that we infer an appropriate rule, especially considering the non-linearity of modern machine-learning techniques, and that is assuming that we get a helpful counterfactual.

One obvious case where one wouldn't receive a helpful counterfactual is when presented with an adversarial example (Ren et al., 2020). Wachter et al. (2018, Sect. II.C) do mention this as a case to avoid (though strictly speaking their definition counts it as an admissible explanation), saying that such cases are outside the "space of real images" and that such artificial inputs should be avoided. Similarly, the definition from Ref. Karimi et al. (2021) requires that the counterfactual cases come from a plausible subspace of possible inputs, given the data. However, natural examples exist too: brightly coloured eyeglasses can trick facial recognition software (Sharif et al., 2016), an ultra-violet spectrum picture of the sun has been interpreted as showing a jellyfish (Hendrycks et al., 2019), as well numerous other examples (Alcorn et al., 2019; Hendrycks et al., 2019). One may want to respond here that such cases will always be distant from the actual input, but even that need not be true. Hendrycks et al. (2019, Fig. 9) present a case where a dragonfly resting on a yellow shovel is classified as a 'banana', but the exact same picture with a red or blue shovel is classified correctly. In contrast, the manipulationist definition I have suggested would either avoid such examples because they are not part of a (relatively simple) generalization that also covers the original input or it would make these cases more interpretable via a broad generalization that covers the causes of such artefacts and when to expect them. I touch on this in Sect. 3 on explanatory depth, but for now I think it's safe to say that the rules one would normally use to explain the output of an algorithm will avoid this problems with a 'counterfactual only' definition. The simple reason: a counterfactual is only helpful when it suggests a reasonable generalization.

Empirical evaluations of XAI methods seems to support this point. Lim et al. (2009) and Lim and Dey (2013) found that giving decision rules that motivate the output of the algorithm (or, alternatively, motivate why some other output did not result) lead to users giving the most accurate predictions of system behaviour. van der Waa et al. (2021) specifically studied the contrasting two different XAI methods—rule-based and example-based—and also seems to support this point. When participants were only presented with two counterfactual cases to each decision (highlighting the most relevant variable) they performed no better on either factor identification or predicting system behaviour than in the situation where no explanations were given. Only having two counterfactual cases, but no rules, appears to be of little help. When presented with a rule (and no counterfactual cases) however, participants scored significantly higher on factor identification, though they were not better at predicting system behaviour. Their hypothesis is that the rules they gave were too narrow, as they only applied to a single decision (as I discuss in Sect. 3, that also implies that they were shallower explanations on the manipulationist definition). To be precise, these rules were counterfactual and of the format "if the alcohol intake would have been 1 unit or less, the system would have advised a normal dose of insulin". Furthermore, Chromik et al. (2021) found that users generalize from a collection of (contrasting, so including counterfactuals) local (Shapley) explanations, and typically do this incorrectly. That further supports that a pure case-by-case approach, where counterfactuals are presented but without overarching generalizations, doesn't truly explain the functioning of an algorithm. Rules might be a better bet, then. That does raise the question: do the rules have to be counterfactual in the way the manipulationists specify? Are other types of rules really incapable of providing explanations? I turn to that question in the next subsection.

2.3 Rules Without Counterfactuals

To structure this discussion I look at the definition of explanation given in Fong and Vedaldi (2017). They hold that

> An explanation is a rule that predicts the response of a black box f to certain inputs. For example, we can explain a behavior of a *robin* classifier by the rule $Q_1(x;f) = \{x \in \mathcal{X}_c \leftrightarrow f(x) = +1\}$,where $\mathcal{X}_c \subset \mathcal{X}$ is the subset of all the robin images. Since f is imperfect, any such rule applies only approximately. We can measure the faithfulness of the explanation as its expected prediction error: $\mathcal{L}_1 = \mathbb{E}[1 - \delta_{Q1(x;f)}]$, where δ_Q is the indicator function of event Q. Note that Q_1 implicitly requires a distribution $p(x)$ over possible images $\mathcal{X}$. Note also that $\mathcal{L}_1$ is simply the expected prediction error of the classifier. Unless we did not know that f was trained as a robin classifier, Q_1 is not very insightful, but it is interpretable since $\mathcal{X}_c$ is (Fong & Vedaldi, 2017, p. 3450).

Is this a helpful definition of an explanation? The issue, I think, is that the rules that adhere to this definition tell you (ideally) *what* a black box does, but do not tell you *why* it gives you those outputs in the specified cases. To stick to the robin example, the explanation proposed here is a set of all robin images. (or more precisely, a set of all images positively classified by the algorithm, but those should mostly be robin images) That tells us something about the classifier, namely that it gives score 1 to precisely those images, but it does not tell us what reasons the classifier has for that classification. Imagine responding to a child that asks ‘why is that a robin?’ after seeing one. The suggested answer here is a long list of robin sightings. Instead, we would probably respond (or can find on Google) ‘because it has a distinctive red breast’, where the ‘distinctive’ part tells you that birds lacking one are given a different name. The interpretability here doesn’t come from the domain, but from the classificatory rule the determines the domain. The definition by Fong and Vedaldi (2017) doesn’t give us such a classificatory rule at all. If that’s not already obvious, compare it again to the baseball example from earlier. Their suggestion is to explain why the window broke with something like the following:

(6) $Q_1(x;f) = \{x \in \mathcal{X}_{breaks} \leftrightarrow f(x) = +1\}$, i.e. we give a long list of cases in which the window breaks.

We might distill from this that the window breaks for $\{v : v > v_2\}$, which is already more informative though still far shallower than more general physical laws. Such a move would, however, seem to interpret the domain over and above what their definition tells us by specifying a rule that fixes $\mathcal{X}_{breaks}$ independently from the function f that we try to explain. (6) specifies, at best, which variables are relevant, but it doesn’t tell us how they influence the output. In other words, we don’t get an answer as to *why* they lead to the outputs they do. The general statement, that explanations predict the response of a black box, is then one I agree with (that is what generalization G does in the manipulationist definition), but their formal specification of the rule differs, tellingly, by not including any counterfactual cases. As a result, there is no independent conceptualization of the domain other than the output of the black box classifier, and so the rules do not yield explanations.

To start applying the manipulationist definition more directly to XAI, I’d like to point out that Fong and Vedaldi (2017) introduce their definition in the context of an improved saliency method that shows which parts of an image are relevant to the classification of it by a black box classifier. What does the above criticism then mean for saliency methods, which visualize the gradient of the classifier in the local neighbourhood of a particular image? First, they differ from the letter of the above definition, as they do not specify a domain where the classifier outputs e.g. ‘robin’, but focus rather on which changes to the image most directly change the output of the image. As a result, saliency maps are actually a closer fit with the ‘counterfactual-only’ definitions from the previous section.[2] Still, they do share an issue in common with the definition of Fong and Vedaldi (2017). Saliency maps show you which pixels/variables are relevant, but they do not tell you *why* they are relevant (empirical evaluations support this claim, e.g. Alqaraawi et al. (2020) found that users struggle to generalize from saliency maps). Similarly, their definition of ‘explanation’ yields an answer that shows you the positively classified inputs, but doesn’t tell you *why* those inputs lead to $f(x) = +1$.

I want to be clear here that my claim that saliency methods do not yield explanations is certainly not meant as a claim that they are not useful. Saliency methods can show us whether a black box attends to the right variables in making its decision. That information alone is valuable: if irrelevant parts of the image (e.g. the sky) show up as highly relevant with saliency methods for making a classification (e.g. whether a tank is visible) then we know that the decision is not made for the right reasons. No ‘proper’ explanation is needed in that case to determine that something is wrong. Furthermore, as with the ‘counterfactual-only’ approach, saliency methods might suggest generalizations to us. If they show that the classifier pays constant attention to fingers when trying to classify a certain fish (e.g. a ‘tench’), it may point us to the realization that it is a fish prized by fishermen and therefore the data shows it almost exclusively when held by a person (as opposed to other fish not shown in this way; Brendel (2019)). Saliency methods can thus be very valuable in diagnosing problems, even if they do not by themselves explain why the highlighted parts of the image lead to the observed output. To be clear on the effect of adopting the manipulationist definition, I briefly discuss in the next subsection what the proposed definition of explanation means for the status of other tools in XAI, using the taxonomy of these methods given by Guidotti et al. (2018).

2.4 Existing XAI Tools and the Manipulationist Definition of Explanation

As I just argued, saliency methods do not provide full (local) explanations of a black box algorithm. So, what is the status of other XAI methods according to the manipulationist definition of explanation? According to Guidotti et al. (2018) these tools can be classified into ones for (i) model explanation, (ii) outcome explanation, (iii) model inspection and (iv) transparent box design. It’ll be clear fairly quickly how the

[2] Fong and Vedaldi (2017) consider saliency methods as local explanations (predicting the response of the classifier in the immediate neighbourhood of a point x_0), whereas the earlier cited definition is for global explanations. Still, saliency methods do not visualize the other outputs in this immediate neighbourhood, but only show which changes to x_0 would alter the classifier outcome the most and so do not seem to fit their local definition either.

definition applies, so I won't discuss the different categories in detail, instead aiming to clarify the effect of adopting a manipulationist definition to whether or not certain tools produce explanations.

Two main methods are discussed under the model explanation heading, namely explanations via single-tree approximation, where the behaviour of the black box is modelled by a single decision tree and explanation via rule-extraction where the behaviour of the black box is modelled by a set of rules. Both cases can readily count as proper explanations under the manipulationist definition, provided the rules (in the decision tree or the set of rules) cover counterfactual cases. This is almost automatically the case for a decision tree, where the alternative children of a node provide counterfactual cases provided the leaf classifications differ from the actual case. With rule extraction counterfactuals might only be supported by considering several rules from the rule set. For example, Craven and Shavlik (1994) describe an algorithm learning M-of-N rules where M features of a list of N features have to be present for a certain classification to be given. This says nothing about what happens when insufficiently many features are present, but one can expect that in practice other M-of-N rules will apply, giving the full set of rules the right features to qualify as an explanation. However, not all methods discussed as model explanations will count as manipulationist explanations. Guidotti et al. (2018, p. 25) also mention, for example, feature importance ranking measures (Sonnenburg et al., 2008; Vidovic et al., 2016). As with the discussion of saliency methods earlier, such tools may not yield full explanations on the proposed definition, but can nevertheless be valuable.

This result, that some of the current tools will not be classified as yielding explanations, is also seen for outcome explanation tools, of which saliency methods are an important part. Such inspection tools can certainly give us further insight into the functioning of a black box algorithm, and help us ensure that they attend to only relevant variables, but they do not on their own explain why a black box reached a specific output. On the other hand, there are also rule-based tools available for outcome explanation, such as LORE (Guidotti et al 2018a), which provides per outcome a decision rule and a set of counterfactual rules that suggests how local changes in input will alter the outcome. LORE does, however, separate the two, so that the counterfactual is not a specific instance of the generalization. For example, the output of LORE can be: rule = $\{age < 26, job = clerk, income = 800, car = no\} \rightarrow deny$ with separate counterfactuals: $\{income > 900\} \rightarrow grant$ and $\{job = employer\} \rightarrow grant$. What happens in the gap between $income = 800$ and $income > 900$? We don't know, because the rule only covers cases with the same output, and does not cover counterfactual instances. As a result, LORE gets close to the definition, but ultimately doesn't fit completely because there is no one rule G that covers both the actual case and a counterfactual instance. Perhaps such a generalization can be reconstructed from the output, but a more explicit specification of one would be more helpful.

The same picture emerges for model inspection tools. Some, such as that presented in Thiagarajan et al. (2016) may seem to get close to yielding explanations as they use a decision tree. The visualization chosen, however, only shows which classes are considered in the different nodes and whether the input in question is classified as such or not (e.g. yes/no on 'is it grass?'). There are no counterfactual cases shown, nor are any generalizations given. One doesn't learn why the model decided that the image doesn't show grass, nor does one learn why e.g. the question 'is it sky?' is considered shortly after. Though it can give us helpful information on the black box, it does not clearly explain why it produces the outputs it does. Other model inspection tools are activation maximization, partial dependence plots and sensitivity analysis for which much the same applies as to saliency methods. It is perfectly possible that one of these visualizations strongly suggests an appropriate generalization, in which case they may lead to explanations (if the suggested generalization is sufficiently accurate), but there is no guarantee that this will happen.

Though it may be disappointing that not all current XAI methods yield explanations according to the manipulationist definition, I think that this is correct. The tools that fail to make the cut are no less helpful for inspecting models and checking their behaviour, but when we look for explanations we have a fairly stringent criterion in mind. They have to answer the question 'why this output?' Pointing to feature importance is a first step to an answer, but not sufficient—hence that such methods fail to make the cut. That being said, there is a last refinement to be made to the discussion on the definition of 'explanation'. Namely, they are not answers to just 'why this output?' but instead are answers to contrastive why-questions. That contrastive aspect of explanation is discussed in the final subsection of the part on defining 'explanation'.

2.5 Contrastive Explanations

As Miller (2019) discusses and formalizes using structural causal models in Miller (2021), plus as is widely argued in the philosophical literature on explanation (Dretske, 1972; Lipton, 2004; Northcott, 2013; Van Fraassen, 1980; Woodward, 2003) explanations are best seen as answers to *contrastive* why-questions. Specifically, explanations are contrastive in both the cause and effect slot (following here the claim that one explains by giving a cause):

X_A rather than X_C explains y_A rather than y_C.

For example, 'the baseball hitting the window at v_1 rather than v_2 explains why the window broke rather than remained intact'. Here X_A, the actual cause, is 'the baseball hit the window at v_1, X_C, the contrastive cause is 'the baseball hit the window at v_2' and the actual and contrastive outcomes are that the window broke/remained intact. All this information can be found in the earlier examples, but I think it helps to make it explicit. There are a few points to make here. First, the fact that this is (arguably) the format all explanations follow, designers of XAI tools can use it to format their answers. Users should be presented with an alternative output (where, as Lipton (2004) argues, 'P rather than not-P' is not a valid contrast) and an input that would lead to it. Perhaps users should be given a range of choices, with different alternative outputs, but the point remains that a contrast should be present. If it isn't, as with saliency methods that do not show alternative input or output, one fails to provide a full explanation of the functioning of an algorithm.

This contrastive element thus gives a concrete format to follow in the design of explanation tools. It is also already incorporated into the definition by Woodward

(2003), though this was perhaps not obvious when I first presented it. He requires that an explanation is given via generalization G where first $G(X_A) = y_A \pm \delta$, i.e. the actual case is covered. And then a contrast is selected, such that $G(X_C) = y_C \pm \delta$. So, if one follows this definition then the explanation will automatically contain a contrast. I still mention the widely agreed upon format of an explanation to underscore that the $G(X_C)$ case is not merely there to ensure that the generalization is an appropriate one (and thus only G needs to be shown, together with $G(X_A) = y_A$), but that the counterfactual case is an important aspect of the explanation itself.[3]

Furthermore, this contrastive format is often not supported by existing XAI methods. Tools such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg et al., 2017) do not allow for the specification of a contrast class. As Watson and Floridi (forthcoming) point out, they often defer to the mean output of the algorithm as the implicit contrast, whereas there may well be other relevant contrasts. So, though the contrastive nature has been noted in the literature already and has received formal treatment in Miller (2021), it bears repeating. That being said, I hope to have made progress on defining 'explanation' in the XAI context. Crucially, though, I haven't said anything yet about what makes for a *good* explanation. That is the topic of the second half of this paper.

3 Explanatory Depth

Granted that the manipulationist definition is a good one to use in XAI, how can we tell which generalizations we should offer to a user? The idea that a wider generalization is better is fairly standard, as I also quoted from Guidotti et al. (2018) in the introduction, but merely offering most general G available does not seem to cut it. Consider the following case: a black box b is explained by $G(X)$, where $G(X) = b(X)$ (i.e. take as generalization the function that the black box approximates). Then it seems that G is the widest generalization available and so should be the most powerful explanation we can give of b. However, the manipulationist definition doesn't require G to have an easily interpretable format such as that of a single decision-tree (and even then, decision trees accurately covering every input of b might be so large as to be hard to follow). It might as well be a point-wise defined function, linking every possible input to the exact output given by b. Is this G, which only links inputs to outputs, the basis for a good explanation? It seems not. There is a good reason that XAI focusses on decision trees and the extraction of relatively short rules, namely that those are generally more helpful than a G that merely replicates the black box.[4] So what does this very wide generalization lack that makes it a less than ideal generalization? I will follow two ideas in this section. First, the abstractness

[3] Furthermore, the contrast likely determines what variables are relevant in the answer, i.e. the contrast determines one's choice of the set X. As such, the contrast also impacts the discussion in Sect. 3.1, as it influences which variables need to be considered at what level of abstraction.

[4] see also (Woodward, 2003, Sect. 5.10) for this type of objection, in his case the ideal gas law that allows you to calculate different cases but whose underlying mechanics were long misunderstood

of the variables used in G (i.e. in any explanation) matters. Second, generality isn't confined to the number of inputs of b, but should also be considered as the number of black boxes to which G can be applied.

3.1 How Abstract Should Variables be?

Which variables we use in our explanations matters. Compare, to start, the following two examples from Woodward (2010) about a pigeon trained to peck at anything that is red:

(7) The pigeon pecked (rather than only looked) because it was presented with a scarlet stimulus (rather than some other stimulus).
(8) The pigeon pecked (rather than only looked) because it was presented with a red stimulus (rather than some other stimulus).

Even though both explanations can be used to explain why the pigeon pecked at a scarlet stimulus, we would prefer (8) over (7). The reason seems to be that 'red' is more general/abstract than 'scarlet' (specifically, every scarlet stimulus is a red stimulus, but not every red stimulus is scarlet). There are two ways to spell this out in more detail. First, we might hold that both (7) and (8) use a variable *colour*, where (7) has the more specific value 'scarlet' and (8) the more abstract/general value 'red'. Second, one can binarize the situation by saying that (7) uses the variable *scarlet*, which has value 1 if the stimulus is scarlet and value 0 if the stimulus has a different colour. (8) then uses the more abstract variable *red*, which has value 1 if the stimulus is red, and 0 if the stimulus has another colour. In both cases we see the difference in abstractness: $colour =$ scarlet implies $colour =$ red, and $scarlet = 1$ implies $red = 1$. I'll opt for the binarized version to avoid confusions between different occurrences of 'colour' and to conform to the standard setup in structural causal models as in Halpern and Pearl (2005a) and Miller (2021). In that case, we'd say that the explanation with rule $G(red) = red$ (the pigeon pecks at red stimuli) is better than the explanation $G(scarlet) = scarlet$ (the pigeon pecks at scarlet stimuli).

This already helps with the black box example above, where b is explained by G point-wise defined on the full set of input variables X. These input variables often have a very low degree of abstraction (e.g. colour values for individual pixels, a single word-string, a sound signal) and that will be one reason for the fact that an explanation which simply presents the function approximated by a black box seems relatively unhelpful. One way to push for more generality in the explanation is to demand more abstract variables (Jansson & Saatsi, 2019). Furthermore, the low level of abstraction is often accompanied by a very large number of input variables to keep track of. Abstraction will, by subsuming variables such as the different shades of red under a single heading, reduce that number and thus make it more cognitively feasible to follow the explanation. The answer seems simply: we should aim for the most abstract variables available. Manipulationists tend do so in terms of the actual value of variables (Blanchard 2020):

> An explanation with explanans variable(s) x_1 is more abstract than an explanation with explanans variable(s) x_2 when the actual value of x_1 is implied by the actual value of x_2, but not vice versa.

(8) gives a more abstract explanation than (7) in this case because *scarlet* = 1 implies *red* = 1, but not vice versa. Similarly, if one wants to explain why an image containing a yellow shovel was classified as ‘banana’ by appeal to the yellow colour of the shovel, the explanation appealing to $x_1 = \textit{yellow shovel}$ (taking value 1 if there is a yellow shovel) is more abstract than $x_2 = \{pixel_1, pixel_2, \dots\}$, because the actual value of *yellow shovel* is implied by the colour values of pixels 1 through n. If, in turn, we appeal not to the colour of the shovel, but to yellow being the dominant colour in the picture (i.e. there being mostly yellow in the image), we see that this gives an even more abstract explanation: *dominant yellow* = 1 is implied here by *yellow shovel* = 1, given that the shovel takes up most of the image. The recent idea to use concept-based explanations (Ghorbani et al., 2019; Yeh et al., 2020) has essentially this same move towards abstractness as they aim for explanations in terms of whether, e.g., the algorithm contains the activation pattern for ‘wheel’ when it predicts ‘car’ instead of looking at individual pixels (as for saliency maps). That abstraction, if one manages to interpret the automatically generated concepts correctly, seems to help. And, indeed, (9) seems to be a better explanation of the classifier’s behaviour than (10):

(9) The classifier says this image contains a banana (rather than a shovel, say) because it mostly contains yellow (rather than some other colour).

(10) The classifier says this image contains a banana (rather than a shovel) because this set of pixels is yellow (rather than some other colour).

The reason that the more abstract explanation is better, so e.g. Woodward (2010) argues, is that it covers more counterfactual situations correctly. Appealing to *red* rather than *scarlet* is better because it more accurately covers cases where the stimulus is a shade of red other than scarlet. (9) is better than (10), presumably, because it seems unlikely that the classifier only considers something a banana when precisely those pixels are yellow. Rather, it is plausible that it also classifies images where a different set of pixels is yellow, provided that yellow is the dominant colour in the image. That gives some basis to the idea that not just any general explanation is a good one, but that it is abstractness in the explanandum that matters for a good explanation. However, as Franklin-Hall (2016) has argued, it isn’t always the case that one should pick the most abstract antecedent possible. For explanation (11) seems less explanatory than (8) even thought the variable is more abstract:

(11) The pigeon pecked because it was presented with something stimulating. Where something is stimulating if it is a red stimulus, or food, or a tickle on the chin, or an electrical signal fed into the cerebellum.

We see that the variable *something stimulating* is more abstract: if *red* = 1, then *something stimulating* = 1, but not vice versa. So (Franklin-Hall (2016) argues) if we are to hold that more abstract variables are always better, then (11) should be a better explanation of the pecking than (8). Furthermore, it is a more complete generalization: it covers all cases where the pigeon pecks (no irrelevant variables/ values are included), so on the generality measure it, too, seems to count as a better explanation (to compare: saying that a scene classification algorithm outputs 'living room' because there are midsized objects is to use an abstract variable, that is relevant and correct, but it's not a very helpful explanation). How do we avoid this consequence? I follow Blanchard (2020) in holding that an explanation that is, aside from more abstract, also more specific is a better one (slightly altered to explicitly include the case where x_1 changes value, also see Woodward (2018) for a response):

> An explanation with explanans variable(s) x_1 is more specific than an explanation with explanans variable(s) x_2 when x_2 is a function of x_1 and other variables $x_3, \ldots x_n$ such that for $x_1 = x_{1,A}$ neither x_1 nor $G(x_2) = G(f(x_1, x_3, \ldots x_n))$ change value if the variables $x_3, \ldots x_n$ are varied.

Basically, the idea is that an explanation is more specific when it doesn't contain variables that are irrelevant to the considered contrast. In (11), the variable *something stimulating* is an example and would count as x_2 here. For, its values are implied by $x_1 = red$, $x_3 = food$, etc (the function here being a Boolean: $x_2 = red \vee food \vee tickle \vee cerebellum$). To see how the definition gives us the result that *red* is more specific, consider the case when $x_1 = red = 1$. As long as the value for *red* is kept fixed, the bird will peck regardless what one does to the values for the other variables. So, $x_2 = something\ stimulating = 1$ in that case, and similarly G(*something stimulating*) = 1. Furthermore, also providing food does not remove the red stimulus that is presented and so doesn't change the value of x_1 (*red* remains 1, and so G(*something stimulating*) still equals 1 even if we change $x_3 = food$). In this case $x_2 = something\ stimulating$ is not specific enough to yield good explanations. Just as $x_1 = midsize\ object$ is not very specific, even though it is abstract. Instead, a variable such as *table* would be better (and more abstract than sub-types of table).[5]

This contrasts with the case of *red* and *scarlet*, where one might see *red* as a disjunction of all the different shades of red. These different shades aren't independent of each other, so if I set $x_3 = bordeaux = 1$, the value of $x_1 = scarlet$ would have to change to 0 (though the bird will still peck as bordeaux is a shade of red). The value of x_1 changes, and so *scarlet* is not more specific than *red* on this definition. That's good news, because the more abstract explanation with *red* instead of *scarlet* is preferred. So specificity doesn't push us all the way back to the least abstract variables, but hopefully settles the matter in an optimal middle ground.

Specificity, then, removes variables from the explanation that aren't relevant to the current case. Though the other variables in (11) will be relevant in other cases (e.g. when only food is presented, so when $red = 0$), it does not give us a good/

[5] Note that we can say the same in terms of values of variables: if we go for the variable *object* with different values, there is again a challenge of saying whether *object* = *midsized*, *object* = *table* or *object* = *dining room table* is a better choice. The same would apply: choose a value that is abstract, but specific.

relevant explanation for the pecking behaviour when only a red stimulus is presented. So, although *something stimulating* is a more abstract variable, it does not lead to better explanations. From the examples discussed here it then seems that a more specific explanation is better than a less specific, more abstract explanation, but that if there is no difference in specificity then we prefer the more abstract explanation. This also nicely fits the idea in Halpern and Pearl (2005a) that explanations should provide minimally sufficient causes (i.e. necessary and sufficient). While that doesn't quite capture the reason to prefer *red* over *scarlet* (we could claim for both that it is necessary and sufficient to present a red/scarlet stimulus for the pecking to occur), it is certainly in the same spirit. There are good reasons, then, to aim for explanations using variables at this level of abstraction. Ideally that claim would get empirical testing, but I think that this account gives a good approximation of the types of variables we prefer to see in our explanations. Abstract, but still specific enough to only mention relevant parts of the actual phenomenon.

3.2 Notions of Generality

Abstractness is one factor to consider, but its relevance, as briefly mentioned in the previous section, seems to come from the fact that more abstract explanations are more general in a specific sense: they correctly apply to more counterfactual cases. Specificity limits that somewhat (explanation (11) correctly applies to even more counterfactual cases than explanation (8) but is worse for it), but clearly generality is important. Hitchcock and Woodward (2003) have, perhaps unsurprisingly, built their account of explanatory depth around generality, and that is the basis I will use here too. According to them a generalization *G* is better if it answers more what-if-things-had-been-different questions. The more invariant a generalization is, i.e. the more can change (to the variables in *G and* background conditions) without dropping below the minimal accuracy, the better the explanation based on it. As Blanchard et al. (2018) argue this falls into two categories. On the one hand, one can consider invariance in terms of the breadth of the generalization, which means the range of cases to which it is taken to apply. For example, the second explanation of why Mary went bungee-jumping has more breadth and is naturally taken to be better:

(12) Mary has gene *g*, which causes people to go bungee-jumping.
(13) Mary has gene *g*, which causes people to engage in risk-taking behaviour.

The generalization underlying (13) will apply to a wider range of cases, and so yields the better explanation. In contrast, there is also the possibility of an explanation being better because it more accurately applies to the cases under consideration. Blanchard et al. (2018) give the following examples:

(14) High levels of cholesterol cause heart disease.
(15) High levels of low-density cholesterol cause heart disease.

In this context, only low-density cholesterol in fact causes heart disease. That means that (14) is wrong in cases where one has high levels of high-density cholesterol (which is assumed to not cause heart disease). Though both will be right in the case of a patient who has low-density cholesterol and got heart disease as a result, still (15) seems the better explanation. The reason is that the generalization in (15) is more accurate than that in (14), as it also gives the right result for high-density cholesterol (namely, that it doesn't cause heart disease). We can translate these two considerations into more formal guidelines for the XAI context.

I'll start with breadth, as this requires a small adjustment to how I have applied the account so far. I wrote primarily about G applying to a single algorithm, explaining why black box b gave output y. One should, however, consider that a generalization may apply to more than one black box algorithm. So, one of the variables in set X for $G(X)$ will stand for the black box to which G is applied. This is a somewhat different generalization than that of (model agnostic) XAI tools which can be used to produce explanations for a wide range of algorithms. The explanations they produce, such as a decision tree, a set of rules, a saliency map or partial dependence plot, generally do not tell you anything about the behaviour of other algorithms. The fact that black box b is described by a specific decision tree does not mean that that decision tree can also be used to explain black box b', even though one may use the same tool to construct a second decision tree for b'.

Instead, I mean that the same explanation can apply to more than one algorithm (something which distinguishes my account from the formal framework of Watson and Floridi (forthcoming)). This does happen when one looks at the informal explanations given. For example, the case where a classifier recognizes the fish species 'tench' based on the presence of fingers. If we ask 'why does the network classify based on fingers rather than aspects of the fish?' then a decent explanation is 'there is a strong correlation between the irrelevant feature (fingers) and the classification goal (tench) in the data, and as a result the easiest way to improve accuracy was to classify based on the irrelevant feature.' This is an explanation that is very broad: it captures the case of tench and can be used to explain not just why we see certain saliency maps, but also why any fish held in ones hands is classified as a tench (i.e. direct output of the algorithm). Furthermore, it applies to a very wide range of black box algorithms. One can use the same generalization to explain why an HR system for engineering jobs auto-rejects applications by women. Of course, an even deeper explanation would tell you exactly how such a correlation affects the output of the black box, but my point is that even this highly simplistic explanation strikes us as a better one than an explanation that makes no mention of the general effect of such correlations in the training data on machine-learning methods. Not because the other models influence the output of the model that is to be explained, but because by giving a more general explanation we can show more thoroughly what causes the output of the model to be explained. How would that output change if we had a slightly different training set? To answer that question we'd, strictly speaking, have to include the different model (because a different input-output function) that results from changing the training set. And yet it's relevant for understanding the outcomes, and even more so for being able to change them. It tells us, for example, that we should include pictures of tench that don't include fingers if we want a more robust classifier.

Such informal explanations have the kind of breadth that a decision tree built to mirror a single black box algorithm does not. They rely on generalization (about spurious correlations in the data, or to give another example, the existence of adversarial examples in neural networks) across algorithms. In that sense I mean to include the black box algorithm as a variable in the set X, as a generalization G that applies to more black boxes will be a better generalization than one that applies to a single black box. This, incidentally, is also a reason why $G(X) = b(X)$ is far from a perfect explanation of a black box algorithm aside from the abstractness of variables. In such a case G lacks the breadth that one of these informal explanations can have (which hopefully we can formalize at some point, when we better understand how neural networks behave). For though it may cover more of the output of b, it does not cover other black box algorithms. A generalization G' that also applies to other black box algorithms will therefore quickly do better on the breadth measure and thus count as a better explanation of the functioning of (part of) b.

Secondly, then there is the question of accuracy. I will spend less time on this, as it is a familiar idea that more accurate generalizations are a better basis for explanations, and it can simply be measured as prediction error made by G in the XAI case. The only change compared to e.g. Fong and Vedaldi (2017) is that strictly speaking more than one black box should be considered. That being said, one could use for example the mean square error with B the range of black boxes to consider and $\mathcal{X}$ the set of possible inputs (though there will be quite a few practical issues in doing so across algorithms):

$$\sum_{b \in B} \sum_{X \in \mathcal{X}} (G(b, X) - b(X))^2 \Bigg/ \sum_{b \in B, X \in \mathcal{X}} 1.$$

The more accurate G is, and the more counterfactual cases it covers, the better an explanation based on G will be. As a reviewer pointed out, however, simply covering more algorithms might not always be a good thing: ideally the algorithms would be relevant for the model to be explained (e.g. variations of the same algorithms but with different training sets, or algorithms of the same type such as convolutional neural networks). Defining that relevance is a challenge I leave to further work. In the mean time, the earlier section showed that abstraction is one way to get to generality, though it has also been argued that abstraction is an explanatory virtue in its own right independent of whether it allows us to answer more what-if-things-had-been-different questions (Weslake, 2010). The discussion so far has also left out arguments for some other explanatory virtues (Hitchcock & Woodward, 2003; Ylikoski & Kuorikoski, 2010) such as cognitive salience and the idea that a more precise contrast (e.g. 'red rather than blue' v.s. 'red rather than any other colour') leads to better explanations. Such aspects may well be important, but are hard to define without strong empirical backing. The idea, on the other hand, that the number of counterfactual situations matters and that from this the relevance of abstraction, breadth and accuracy follows, is more clearly motivated by the manipulationist definition of explanation which holds that explanations answer what-if-things-had-been-different questions. Though it is likely to be a tough challenge to design tools that yield excellent explanations on this framework, I do hope that this definition and the considerations around explanatory depth can help XAI by setting clearer goals than have so far been available.

4 Conclusions

What does it mean to explain the outcome of an algorithm? I have presented the manipulationist definition of explanation, and how it applies in the XAI context. The answer, briefly put, is that it means answering what-if-things-had-been-different questions and that one does so by giving a generalization G that covers the actual case and at least one counterfactual case. This ties together existing definitions that focussed on either only counterfactuals (e.g. algorithmic recourse) or only on rules without care for counterfactuals (Fong & Vedaldi, 2017). Furthermore, it leads to a natural definition of when one explanation is better than another: if more what-if-things-had-been-different questions are answered, the explanation is better. This can be done by employing more abstract variables, using generalizations with more breadth (possible spanning more than one algorithm) and using more accurate generalizations. Still, it will likely be hard to do so in practice with the tools currently in place. So what can one practically do with the account presented here? A few of the more practical upshots of the definitions discussed here are:

- Present explanations in the contrastive format: X_A rather than X_C explains why y_A rather than y_C is the output of the algorithm.
- Include a generalization in the explanation, rather than just one or two counterfactual cases (or, vice versa, a rule without a contrast case).
- When offering explanations, consider the abstractness of the variables in the explanandum X.
- Not only the accuracy of an explanation matters, but its breadth too; a single broader generalization might yield a better explanation than a large set of narrow generalizations. A smaller decision tree might be more explanatory than a larger, and somewhat more accurate, one.

These are only theoretical guidelines, of course, and empirical evaluation (e.g. in terms of how well users can predict the output of an algorithm, and subsequently if they manage to only act on the output if it is correct) such as in van der Waa et al. (2021) will be valuable. By designing explanations by hand the current definitions can be verified, where at minimum they should improve the ability to predict algorithm output. In addition, that empirical work will likely spur the need for the explicit inclusion of pragmatic aspects of explanation, and of modelling the interaction between the questioner and answerer. I haven't included those elements here, to keep the focus on the underlying logic of explanation that (I've argued) remains the same despite the eventual inclusion of elements such as background knowledge and the dynamic setting in which XAI tools will operate. In short, there is plenty of work left to do, but hopefully this theoretical framework makes it clearer what that work is.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Alcorn, M., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W., & Nguyen, A. (2019). Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4845–4854).

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In *IUI '20: Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 275–285).

Blanchard, T. (2020). Explanatory abstraction and the goldilocks problem: Interventionism gets things just right. *The British Journal for the Philosophy of Science, 71*(2), 633–663.

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2018). Stability, breadth and guidance. *Philosophical Studies, 175*, 2263–2283.

Brendel, W. (2019). Neural Networks seem to follow a puzzlingly simple strategy to classify images. *Medium*. Retrieved from https://medium.com/bethgelab/neural-networks-seem-to-follow-a-puzzlingly-simple-strategy-to-classify-images-f4229317261f

Chromik, M., Eiband, M., Buchner, F., Krü ger, A., & Butz, A. (2021). I think I get your point, AI! The illusion of explanatory depth in explainable AI. In *IUI '21: 26th International Conference on Intelligent User Interfaces* (pp. 307–317).

Ciatto, G., Schumacher, M., Omicini, A. & Calvaresi, D. (2020). Agent-based explanations in AI: Towards an abstract framework. In D. Calvaresi et al. (Eds.) *Explainable, transparent autonomous agents and multi-agent systems 2020, lecture notes in artificial intelligence*, Vol. 12175 (pp. 3–20).

Craven, M., & Shavlik, J. (1994). Using sampling and queries to extract rules from trained neural networks. *Machine Learning Proceedings, 1994*, 37–45.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. Preprint retrieved from http://arxiv.org/abs/2006.11371

Dretske, F. (1972). Contrastive statements. *Philosophical Review, 81*(4), 411–437.

Fong, R., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE international conference on computer vision (ICCY), Venice, Italy, 2017* (pp. 3449–3457).

Franklin-Hall, L. (2016). High-level explanation and the interventionist's 'variables problem'. *The British Journal for the Philosophy of Science, 67*(2), 553–577.

Ghorbani, A., Wexler, J., Zou, J., & Kim, B. (2019). Towards automatic concept-based explanations. Preprint retrieved from http://arxiv.org/abs/1902.03129

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018a). Local rule-based explanations of black box decision systems. Preprint retrieved from http://arxiv.org/abs/1805.10820

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42.

Halpern, J., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science, 56*(4), 843–887.

Halpern, J., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science, 56*(4), 889–911.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2019). Natural adversarial examples. Preprint retrieved from http://arxiv.org/abs/1907.07174

Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, Part II: Plumbing explanatory depth. *Noûs, 37*(2), 181–199.

Jansson, L., & Saatsi, J. (2019). Explanatory abstractions. *The British Journal for the Philosophy of Science, 70*(3), 817–844.

Karimi, A., Barthe, G., Schölkopf, B., & Valera, I. (2021). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. Preprint retrieved from http://arxiv.org/abs/2010.04050

Kenny, E., & Keane, M. (2021). *On generating plausible counterfactual and semi-factual explanations for deep learning. AAAI-21* (pp. 11575–11585).

Lim, B., & Dey, A. (2013) Evaluating intelligibility usage and usefulness in a context-aware application. In M. Kurosu (Ed.) *Human-computer interaction. Towards intelligent and implicit interaction. HCI 2013. Lecture notes in computer science*, Vol. 8008 (92–101).

Lim, B., Dey, A., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *CHI '09: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2119–2128).

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). Routledge.

Lundberg, S., & Lee, S. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in neural information processing systems, Vol. 30 (pp. 4765–4774).

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–39.

Miller, T. (2021). Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review, 36*, E14.

Northcott, R. (2013). Degree of explanation. *Synthese, 190*, 3087–3105.

Pearl, J., & Mackenzie, D. (2019). *The book of why: The new science of cause and effect*. Penguin.

Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering, 6*(3), 346–360.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Rosenfeld, A., & Richardson, A. (2019). Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems, 33*, 673–705.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Sharif M., Bhagavatula S., Bauer L., Reiter, M. (2016). Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, Vienna, Austria* (pp. 1528–1540).

Sonnenburg, S., Zien, A., Philips, P., & Rätsch, G. (2008). POIMs: Positional oligomer importance matrices—Understanding support vector machine-based signal detectors. *Bioinformatics, 24*(13), i6–i14.

Thiagarajan, J., Kailkhura, B., Sattigeri, P., & Ramamurthy, K. (2016). Tree- View: Peeking into deep neural networks via feature-space partitioning. Preprint retrieved from http://arxiv.org/abs/1611.07429

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291*, 103404.

Van Fraassen, B. (1980). *The scientific image*. Oxford University Press.

Vidovic, M., Görnitz, N. Müller, K. & Kloft, M. (2016). Feature importance measure for non-linear learning algorithms. Preprint retrieved from http://arxiv.org/abs/1611.07567

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

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

Weslake, B. (2010). Explanatory depth. *Philosophy of Science, 77*, 273–294.

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

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy, 25*, 287–318.

Woodward, J. (2018). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese, 198*, 237–265.

Yeh, C., Kim, B., Arik, S., Li, C., Pfister, T., & Ravikumar, P. (2020). On completeness-aware concept-based explanations in deep neural networks. Preprint retrieved from http://arxiv.org/abs/1910.07969

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies, 148*, 201–219.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.