Two Types of Explainability for Machine Learning Models

Word count: 4993

Abstract

This paper argues that there are two different types of causes that we can wish to understand when we talk about wanting machine learning models to be explainable. The first are causes in the *features* that a model uses to make its predictions. The second are causes in the *world* that have enabled those features to carry out the model's predictive function. I argue that this difference should be seen as giving rise to two distinct types of explanation and explainability and show how the proposed distinction proves useful in a number of applications.

1 Introduction

The recent call for explainability in machine learning is typically spoken of in terms of the need to *explain a model's outputs*. Why has a particular model, f, classified Alice as possessing rare disease y? Why has g classified Bob as ineligible for a bank loan? Why has h predicted that patients $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ possess a high risk of dying from pneumonia whilst predicting that patients $\boldsymbol{x}_k, \ldots, \boldsymbol{x}_n$ possess a low risk? Having the ability to answer such questions appears to be of paramount importance. Yet, what exactly are we asking for and what kinds of methods do we need to get us the answers we want?

Whilst most authors agree that what we are asking for with such questions is knowledge of causes (Watson 2020; Barocas et al. 2020), the question of *which* causes we want to understand does not currently have a clear answer. In this paper, I argue that there are, in fact, two distinct kinds of causal inquiry that we can have in mind when we ask questions of the form *why did f classify $\boldsymbol{x}$ as y*. On the one hand, we can seek an understanding of the causal relationships that hold between the features in a model's input and that model's output values. On the other hand, we can seek an understanding of the causal relationships in the world that have enabled a model to use those particular features to carry out its predictive function. I argue that these different types of causal inquiry are best understood as capturing two distinct types of explainability: *feature-oriented explainability*, and *world-oriented explainability* and show how drawing the distinction proves useful in answering two pressing questions: 1) do the explanation methods emerging from the field of explainable AI (xAI) provide the level of understanding required to achieve fair, accountable and safe machine learning? And, 2) should we accept decreases in model accuracy for increases in explainability?

The paper is structured as follows. Section 2 introduces some technical background. Section 3 argues that there are two kinds of explanations that we can want in machine learning. Section 4 addresses three objections. Section 5 spells out two positive implications of the proposed distinction. Section 6 concludes.

2 Preliminaries

2.1 Machine Learning Terminology and Notation

I shall refer to ML models throughout by way of standard function letters f, g, and h. I represent vectors using boldface italics and scalars with non-boldface italics. When it would be otherwise unclear from the context, I use subscripts to represent particular inputs, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_k$ would thus represent two different vector inputs. Bracketed superscripts are used to individuate particular vector components, where $x_i^{(j)}$ would thus be the jth component in the input vector $\boldsymbol{x}_i$.

Unlike classical AI models, the mapping from inputs to outputs performed by a ML models is acquired through an automated learning procedure or *training phase*. During this phase, a model learns to transform inputs to outputs by exploiting correlations between *features* present within its inputs and the pre-specified output values associated with those inputs. The nature of these features will depend on the model class. The features used in a linear regression model will simply be the individual variables that make up the model's inputs, values of which are represented as individual vector components $x^{(1)}$, ... ,$x^{(d)}$. The features in a image classifier, by contrast, will be emergent patterns *within* these individual input variables. Hence, an image classifier

trained to classify images of animals might learn to exploit the feature *doggy-nose* as a predictor of the class 'dog,' where this feature supervenes upon a collection of more basic input variables (one variable for each pixel). After training, we say that a model has been 'fit' to the data.

2.2 Explainable AI

In this paper, 'explainable AI', or 'xAI' will refer to the loose sub-field in computer science focused upon developing methods to explain the outputs of machine learning models. Whilst these methods differ greatly in their mathematical details there are two straightforward generalisations that we can make. The first concerns the nature of the questions that these methods seek to answer (Watson 2020, 5). These questions take the following form.

> "Q: Why did model f predict outcome y_a as opposed to alternative $y_c \neq y_a$ for some input vector $\boldsymbol{x}$?"

Here y_a is a the actual value observed for the output variable Y, whilst y_c is a contrast (counterfactual) value for Y.

The second generalisation we can make is that these methods seek to identify *causes* and, in particular, a certain kind of cause. No matter the method, the explanation methods in xAI seek to discover which of an input $\boldsymbol{x}$'s features caused the output variable Y to take its actual value y_a rather than some counterfactual value y_c. To take the above example, we might use a particular explanation method and discover that it is the presence rather than the absence of the *doggy-nose* feature that is causing our image

classifier to classify inputs as dogs rather than cats. [1]

2.3 Causal Explanation

I make the assumption in this paper that the explanations sought in calls for explainability are *singular causal explanations*, where I shall assume a contrastive counterfactual account of singular causal explanation. More specially, following several other contemporary accounts, I shall assume that singular causal explanations are i) relativised to a set of background conditions, and ii) contrastive in both cause and effect slots (Northcott 2015; Van Fraassen 1980; Woodward 2003). That is, I will assume that singular causal explanations take the following form.

> Given b_a, C's taking the value c_a rather than c_c explains E's taking the value e_a rather than e_c.

Here C is the explanans variable, E is the explanandum variable, and c_a/e_a and c_c/e_c are actual and counterfactual values of C/E, respectively. b_a represents the actual background conditions that held in the circumstance of interest.

To make this concrete, we might have:

MATCH1: Given the presence of oxygen in the atmosphere (b_a), the match's being

[1] Whilst some xAI methods employ explicit causal discovery methods (e.g: Chalupka et al. (2015)), most instead probe correlations holding between features and output values as a guide to identifying causes. Issues concerning these methods are beyond the scope of this paper but see Hooker & Mentch (2019) and Barocas et al. (2020) for critical commentary.

struck (c_a) rather than not struck (c_c) explains its lighting (e_a) rather than not lighting (e_c).

Note that whilst the presence of oxygen is acting as the background condition or support factor, b_a, in this explanation, we could switch our focus to the presence of oxygen *itself* as the explanans. This would give us:

MATCH2: Given the match's being struck (b_a), the presence (c_a) rather than absence (c_c) of oxygen in the atmosphere explains the match's lighting (e_a) rather than not lighting (e_c).

In MATCH1, we are essentially taking the presence of oxygen as fixed and ascribing causal responsibility for the effect to the match's being struck. In MATCH2, by contrast, we flip this around. Now we are treating the match's being struck as fixed and ascribing responsibility to the presence of oxygen.

Different explanations of the same explanandum may arise like this when the interests of different parties pull in different directions. To take a non-trivial example, when there was a fatal gas leak at a pesticide plant in Bhopal, India in 1984, the pesticide company, Union Carbide, claimed that it was human interference that explained the leak. An independent group of investigators, on the other hand, claimed that it was lax safety measures and decaying facilities that explained the leak.[2] Whilst these latter investigators were not disputing the causal relevance of human interference, they choose to pick out the background conditions or 'systemic causes' to focus on in

[2]See Eckerman (2005)

their explanation (Hanley 2021).

3 Two Types of Explainability

I now claim the following: when we ask why model f classified $\boldsymbol{x}$ as y_a rather than y_c there are two kinds of explanation that we can want. One is an explanation in terms of $\boldsymbol{x}$'s features; the other is an explanation in terms of the background conditions that *enabled* those features to play a causal role. 3.1 backs up this claim. 3.2 presents the distinction more formally.

3.1 Motivating the Distinction - Two Cases

LOAN

You are given a model, g, trained to evaluate the credit-worthiness of loan applicants. g has been trained to map applicants, represented as vectors $\boldsymbol{x} \in R^d$ to a credit-worthiness (CW) score, where each of the vector components in an input $\boldsymbol{x}_i$ represents the value taken by a different feature–*salary*, *education*, *past credit history*, etc–in that particular applicant. Now, suppose g classifies an applicant, Bob, as having a low CW score. You subsequently deny Bob a loan. Bob wants to know why the model has classified him in this way so that he can attempt to make the changes required to receive a more favourable outcome when he reapplies in the future. He is asking for an *explanation* of the model's output, but what exactly does this mean? What exactly is it that Bob wants to understand in this context?

HOSPITAL

You have been tasked with reforming hospital admissions policy for pneumonia patients. You are given a neural network model, h, trained on a corpus of hospital admissions data. The model has learned to predict probability of death (POD) for patients diagnosed with pneumonia using a set of d patient features: *age*, *sex*, *routine blood pressure*, *white blood cell count*, etc. You want to use h to help you allocate scarce hospital resources to those most in need. In particular, you would like to treat those pneumonia patients with a low POD score as outpatients and admit only those patients whose POD lies above some particular threshold. In order to do so, however, you first need to trust that your model can provide you with information that you can act on. For this, you need to know why the model is producing the particular outputs that it is. That is, you need to be able to *explain* h's *outputs*. But, again, what does this mean? What exactly do you need to know, or understand, in this context in order to be able to trust your model for the purposes of altering admissions policy?

What are the answers? Well, I take it that Bob wants to know which feature values caused the model to classify him as having a low CW score. More precisely, Bob wants something like the following: 'The fact that you have an income of $income_a$ rather than $income_c$ explains f's classifying you with a CW score's lower than our threshold for credit.'[3] What about HOSPITAL? Here I suggest that you need to know not only *how* certain features are being used to make predictions of mortality risk but also *why* those

[3]Of course, in order to attribute the entirety of the difference in CW here to the value taken by the feature *income*, one has to treat all the other values taken by features of $\boldsymbol{x}$ as part of the stable background conditions. I omit this detail to simplify discussion.

features are being used. Let me spell out what this means.

In the paper from which HOSPITAL was adapted, the model was found to be using the feature *has asthma* as a predictor for low *POD* (Caruana et al. 2015). Let's suppose this holds in our toy case too. Now, let's suppose that an asthmatic patient, $\boldsymbol{x}$, is diagnosed with pneumonia and your model classifies her as having a *low POD*. You ask: why has h classified input $\boldsymbol{x}$ as having a low rather than, say, middling *POD*? Consider the following explanation.

> HE1: Given the causal relationships present in the system that generated h's training set $\mathcal{D}$, the fact that $\boldsymbol{x}$ possesses the value 1[4] for the feature *has asthma* explains why h classified $\boldsymbol{x}$ as having a low rather than middling *POD*.

Like the explanation given to Bob in LOAN, HE1 cites the value taken by a particular feature in the input $\boldsymbol{x}$. Yet, this kind of explanation is now inadequate for the task of assessing whether h can be safely used for the purposes of reforming hospital admissions policy. For this, you need to know *why* the model is using *has asthma* as a predictor for low *POD* and this requires understanding the causal relationships in the system that has generated h's training data. For instance, in the example from which the case was adapted, the correlation between *has asthma* and low *POD* was due to the fact that asthmatics in the hospitals from which the data was generated received much

However, it necessarily complicates the actual practice of delivering explanations of this nature. For one thing, the party doing the explaining has to choose which feature to pick out as the explanatory one, something they must do by first considering the needs and interests of the agent requesting an explanation (SeeBarocas et al. (2019) for discussion.

[4]Where: 1='Yes'; 0='No.'

more *aggressive* pneumonia treatment (Caruana et al. 2015). If you used the explanation HE1 as your guide in reforming admissions policy, you might simply choose to let asthmatic patients diagnosed with pneumonia be treated as outpatients. After all, being asthmatic is a predictor of low mortality risk. But this would obviously be a fatal mistake. The kind of explanation we have in mind when we talk about wanting to be able to trust a model to guide safe interventions cannot, therefore, be the same as that provided for cases like LOAN. What I suggest we want instead in this case is something like the following.

> HE2: Given that $\boldsymbol{x}$ includes the value 1 for the feature *has asthma*, it is the fact that asthmatic patients in the datasets from which h was trained received much more aggressive treatment that explains why h classified $\boldsymbol{x}$ as having a low rather than middling POD.[5]

As these two cases have hopefully shown, there are two different kinds of causal stories that we can want in the context of machine learning. We can want a causal story that concerns how features in an input are causing certain output behaviour. Or, we can want a causal story that concerns why those particular features are causally relevant in the first place. With this on the table, let me go ahead and spell out my proposal my formally.

[5]Notice that treating the training environment as fixed in HE1 allows us to remain ignorant about the influence of this cause on the explanandum, whilst treating it as a difference-maker in HE2 demands we understand its causal role.

3.2 Feature-oriented and World-oriented Explainability

I propose we distinguish between two types of explanation for ML model outputs.

Feature-oriented (FO)- explanation:

A FO-explanation for f's producing output y_a rather than y_c in response to an input $\boldsymbol{x}$ is a true account of how the value taken by a certain feature present within $\boldsymbol{x}$ made the difference between f's classifying $\boldsymbol{x}$ as y_a rather than y_c). (Call this explanatory feature $\mathcal{F}$.)

World-oriented (WO)-explanation:

A WO-explanation for f's producing output y_a rather than y_c in response to an input $\boldsymbol{x}$ is a true account of the factors present in the world that enabled the explanatory feature, $\mathcal{F}$, to make the difference between f's classifying $\boldsymbol{x}$ as y_a rather than y_c.

This naturally leads to the following distinction between two different types of explainability, where explainability is i) predicated of ML models, and ii) relativised to an agent or inquirer.

Feature-oriented (FO)-explainability:

A model f is FO-explainable relative to an agent S to the degree to which S can provide FO-explanations for the values taken by f's outputs.

World-oriented (WO)-explainability:

A model f is WO explainable relative to an agent S to the degree to which S can

provide WO-explanations for the values taken by f's outputs.

With the proposal now fully on the table, below I turn to three objections that one might raise.

4 Objections

Objection 1: I am unconvinced by your move to call these different 'types' of explanations. One does not call an explanation that cites the presence of oxygen as the explanans a different type of explanation to that which cites the match's being struck.

Whilst we might not typically refer to an explanation that focuses on background conditions as a different kind of explanation to one that focusses on proximal or 'triggering' causes, there are contexts for which it seems both appropriate and useful to talk this way. Take a case like that of the Bhopal gas tragedy. There were solid reasons why one should care more about identifying the background conditions or 'systemic causes' here than identifying the specific details of the proximal cause. Identifying lax safety measures at the plant, for instance, provided information about a powerful variable which could be intervened upon to prevent similar future disasters (Hanley 2021). Given the asymmetry in importance between knowledge of background conditions and knowledge of proximal causes in this case, it seems perfectly appropriate to say that Union Carbide had, in fact, provided the wrong *kind* of explanation when they explained the disaster solely in terms of proximal causes. Given that it seems not only permissible but *useful* to speak in this way, I see no reason for rejecting the distinction I draw

between FO- and WO-explanations on these grounds.

> **Objection 2:** The distinction between types of explainability you argue for can be captured just as well by the distinction between *partial* and *complete* explainability. FO-explainability is partial in the sense that it only involves being able to identify proximal causes, whilst WO-explainability is complete in the sense that it involves being able to identify proximal causes *and* background conditions.

Whilst perhaps initially appealing, the distinction between partial and complete explainability turns out to be too crude to capture the distinction of interest here. In particular, there are good reasons for wanting to be able to distinguish between degrees of FO-explainability itself. Indeed, this is why I defined these concepts as degree concepts to begin with. Simply calling FO-explainability 'partial' removes our ability to do this. Let me explain.

Machine learning models differ greatly in their mathematical complexity.[6] Whilst some ML models are so simple that one is able to fully anticipate how changes to a model's features will yield changes to its outputs[7], others are so complex that the causal relations between features and outputs will be utterly mysterious. The entire enterprise of xAI is effectively an attempt to take models of the second kind and move them closer to models of the first kind; to 'open up the black box'. Now, being able to assess the degree to which such methods success in doing this is important, but it crucially relies on us being able to talk about different degrees of (FO) explainability. By accepting the

[6]Differences in model complexity primarily come through differences in the dimensionality of a model's inputs.

[7]Such models are often referred to as 'glass boxes.'

proposal in objection 2 and referring to all FO-explainability as mere 'partial explainability,'however, we effectively lose the ability to do this. This is no good. Better to distinguish between FO- and WO-explainability and buy ourselves the expressive power to talk about FO-explainability itself as coming in degrees.

> **Objection 3:** Your analysis is too simple. There are far more things we want to understand than just these two types of causes when we talk of wanting explainability in machine learning.

This may well be true. 'Explainability' is, like 'objectivity,' a highly fuzzy notion used in many different ways. One further way 'explainability' gets used is as a way to refer to the semantic intelligibility of a model's features. Hence, a model might be said to be explainable in this sense if there is there exists some mapping between its space of features and our space of (human) concepts. This certainly deviates from the meaning of both FO-explainability and WO-explainability. Nevertheless, even if there are more senses of explainability that we want to explicate, this fact alone does not detract from the value of distinguishing the two kinds of explainability that I have identified in this paper.

5 Implications

5.1 Putting xAI in its place

In section 2 I introduced the methods in xAI as concerned with discovering a particular class of causal relationships, namely, those that hold between a model's input features

and its output values. The conceptual proposal made in this paper now allows us to say that these methods are engaged in delivering a specific *type* of explainability: FO-explainability. Giving the type of explainability provided by xAI methods a name in this way turns out to be very useful in clarifying what these methods can and cannot do. Let me explain.

What follows is an argument rarely stated but implicitly suggested by many in the recent literature.

P1. The causal understanding entailed by explainability is sufficient to achieve a particular package of moral and epistemic goods.
P2. The technical tools being developed in xAI are sufficient for achieving explainability.

C. The technical tools being developed in xAI are sufficient to achieving that particular package of moral and epistemic goods.

Here these 'moral and epistemic goods' include the ability to i) justify a model's decisions, ii) audit a model for bias and discrimination, and iii) spot when a model is making use of accidental or potentially dangerous features.

The optimistic conclusion, C, is perhaps seen most clearly in the AI Ethics literature where authors typically work with a loose picture of how the methods in xAI work (e.g: Maclure 2021). However, it can also be seen in the various ways in which computer scientists sell the benefits of their own methods (Pavlus 2016; Caruana et al. 2015). The problem, of course, is that none of these goods can be achieved using the understanding

provided by xAI methods alone. Rather, each of these goods requires an understanding of *why* certain features are being used and, whilst one may be able to move from a FO-explanation*to* a WO-explanation, that move is by no means guaranteed. C is, therefore, false.

Consider what would be required in order to defend the use of a model as just. Whilst one would certainly need to understand which features the model was using, the task of justification also requires that one know *why* the model had been able to use those features. Suppose, for instance, that a credit approval model is found to be rejecting loan applicants on the basis of some highly obscure feature, say feature X^{76}. Whilst you have an FO-explanation for the model's behaviour here, you have no idea what this feature is tracking in the real world; for the purpose of justification, this is a big problem. Or, consider a team wishing to investigate whether an email spam filter unfairly discriminates against certain users. Discovering that a particular set of features (words) are highly diagnostic of spam does not by itself tell you anything about whether the model is biased against certain users. Rooting out bias in this way requires that you understand the causal structure of the system that has generated the data upon which the model has been trained.[8] Whilst FO-explanations might provide clues as to what this structure is, there is no direct route from FO-explainability to WO-explainability. Or, finally, consider the desire to identify when a model has learnt to exploit spurious or accidental features. Been Kim sells her TCAV method as being able to do exactly this (Pavlus 2016). Yet, notice something important. To say that a feature is accidental or spurious is to say that the model is using that feature for reasons that have nothing to do with the underlying system that you are trying to predict. But this means identifying

[8]See Burrell (2016) for discussion of the demands of bias detection in this context.

spurious features requires an understanding *of* the system that one is trying to predict! Such understanding is not provided by simple FO-explanation; rather, you need WO-explanation; and, once again, there is no direct route from the former to the latter.

Now, one might worry here that I am making too big an issue of the epistemic gap between the kind of understanding afforded by xAI methods and the kind of understanding required in order to achieve these moral and epistemic goods. But this worry would miss the point here. The point is not that the gap is necessarily unbridgeable or even that the existence of such a gap is particularly threatening in all contexts. The crucial point is simply that the epistemic gap exists and that there are good practical reasons for wanting to draw attention to it. Distinguishing between FO- and WO-explainability as I propose has the attractive consequence of doing just this.

5.2 Re-thinking the Accuracy/Explainability Trade-off

Finally, distinguishing between two types of explainability has one straightforward implication for current debates around the so called accuracy-explainability *trade-off* in machine learning: it forces us to be more precise about what we are actually talking about. Are we comparing the value of accuracy against the value of FO-explainability, or of WO-explainability? Interestingly, clarifying our language in this way turns out to do real work in helping us think more clearly about the nature of the trade-off.

Consider a context for which the accuracy-explainability trade-off appears pressing: clinical diagnosis. One of the key motivations for wanting explainability in this context is the thought that patients should be able to know why they received the particular diagnosis that they did.[9] Yet, I take it that the type of explainability desired in this

[9]Maclure (2021) even argues that doctors should have a legal duty to provide such

context is best thought of not as FO-explainability but rather as the thicker and much more demanding, WO-explainability. After all, FO-explanations will typically be quite meaningless to a patient. Telling a patient that a model has diagnosed him/her with a particular rare disease because *resting heart rate* took a particular value is quite obviously to stop short of telling them what they actually care about, namely, why this feature is diagnostic of the disease in the first place.[10] Since what a patient wants here is thus presumably WO-explainability, let us suppose for the sake of argument that our preferences for FO-explainability, WO-explainability and accuracy in the context of clinical diagnosis are ranked as follows.

$$\text{WO-explainability} > \text{accuracy} > \text{FO-explainability}$$

Whilst I do not wish to pass judgement on how the trade-off should be resolved in this context, it is at least highly significant that separating out the two senses of explainability provides us with a richer representation of our preferences. This may indeed help resolve the issues. For instance, if WO-explainability turns out to be largely out of reach in the diagnostic context due to our limited understanding of the relevant disease mechanisms (London 2020), then maybe we should conclude to resolve the trade-off in favour of *accuracy*. Again, I do not wish to suggest that this conclusion is correct; rather, I aim merely to highlight how distinguishing between the two types of explainability can help us to think more clearly about what it is we are actually trading off in such cases. This is another positive consequence of my proposal.

explanations.

[10]See Watson (2020) for similar discussion.

6 Conclusion

This paper has argued that calls for explainability can be fruitfully understood as calls for two different types of causal understanding. Accordingly, I have proposed that we distinguish between two distinct types of explainability for machine learning models. Whilst there may well be further senses of explainability that we wish to identify, distinguishing between explainability that is *feature-oriented* and explainability that is *world-oriented* is an important step towards getting a better grip on the concept and understanding the different methodological demands that come with its use.

References

Barocas, Solon, Andrew D. Selbst, and Manish Raghavan. 2020. "The hidden assumptions behind counterfactual explanations and principal reasons." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 80–89.

Burrell, Jenna. 2016. "How the machine 'thinks': understanding opacity in machine learning algorithms" *Big Data Society* 3: 1.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "Intelligible Models for Healthcare." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 1721–30.

Chalupka, Krzysztof, Pietro Perona, and Frederick Eberhardt. 2015. "Visual Causal

Feature Learning." *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 181-90.

Eckerman, Ingrid. 2005. *The Bhopal Saga—Causes and Consequences of the World's Largest Industrial Disaster.* India: Universities Press.

Hanley, Brian J. 2021. "What Caused the Bhopal Gas Tragedy? The Philosophical Importance of Causal and Pragmatic Details." *Philosophy of Science* 88: 4, 616-37.

Hooker, Giles and Lucas Mentch. 2029. "Please Stop Permuting Features: An Explanation and Alternatives. arXiv:1905.03151v1 [stat.ME]. https://arxiv.org/abs/1905.03151

London, Alex. 2020 "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability." *Hastings Center Report* 49: 1, 15-21.

Maclure, Jocelyn. 2021. "AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind." *Minds Machines* 31: 421–38.

Northcott, Robert. 2015. "Degree of explanation." *Synthese* 190: 15, 3087-105.

Pavlus, John. 2019. "New Approach to Understanding How Machines Thinks." *Quanta.* https://www.quantamagazine.org/been-kim-is-building-a-translator-for-artificial-intelligence-20190110/.

Van Fraassen, Bas. 1980. *The scientific image.* Oxford: Oxford University Press.

Watson, David S. 2020. "Conceptual challenges for interpretable machine learning." *Synthese* 200: 65.

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

Zerilli, John. 2022. "Explaining Machine Learning Decisions." *Philosophy of Science* 89:1-19.