Causal and Non-Causal Explanations of Artificial Intelligence

Christopher Grimsley

4,818 words

Abstract

Deep neural networks (DNNs), a particularly effective type of artificial intelligence, currently lack a scientific explanation. The philosophy of science is uniquely equipped to handle this problem. Computer science has attempted, unsuccessfully, to explain DNNs. I review these contributions, then identify shortcomings in their approaches. The complexity of DNNs prohibits the articulation of relevant causal relationships between their parts, and as a result causal explanations fail. I show that many non-causal accounts, though more promising, also fail to explain AI. This highlights a problem with existing accounts of scientific explanation rather than with AI or DNNs.

1 The Need for Explainable Artificial Intelligence

The use of artificial intelligence (AI) has expanded considerably in the past decade. AI is increasingly being used to make high-stakes decisions, often under questionable circumstances that indicate the presence of racial or gender bias, including granting or denying loan applications (Fuster et al. 2018), deciding which prisoners are eligible for parole (Khademi and Honavar 2019), and diagnosing mental health disorders (Bennett et al. 2019). If AI is used to make these decisions — especially if these decisions appear to have reinforced biases present elsewhere in society — understanding how the algorithm made the decision is essential. Absent explanation, arbitrary or biased decisions may go unchecked. Computer scientists have recognized this problem and have begun developing explainable AI (XAI), but many of their strategies haphazardly employ a mix of causal, psychological, and counterfactual strategies that fail to generate adequate explanations. It is impossible to explain AI without first explaining explanation. The philosophy of science is uniquely positioned to take on this problem and offer solutions by examining the meaning of scientific explanation and developing an account of explanation which adequately explains AI.

An explainable algorithm is one for which a true, satisfactory explanation exists. An interpretable algorithm is one for which a complete account of the relationships between the steps in the algorithm exists. In many cases, AI decision and classification algorithms are neither explainable nor interpretable. Many of the AI algorithms used in these cases are deep neural networks (DNNs), a type of algorithm whose complexity defies explanation in a particularly striking manner. Because explanation through merely technological means is lagging behind the complexity of the networks that are in need of an explanation, it is reasonable to conclude that the solution to this problem cannot be technological. If this is the case, a potential solution can be found in the ways in which explanation is conceptualized within the context of AI. In order to solve the explainability problem, it is first necessary to articulate an appropriate model of explanation which can be effectively applied in this context.

I argue that recent attempts by computer scientists to develop XAI fail because they do not employ a theoretically-grounded concept of explanation. Further, I show that it is necessary to employ non-causal accounts of explanation in order to solve the problem of explainability in AI. I begin with a brief overview of the aspects of AI that are relevant to my argument. Then I discuss two existing methods for developing XAI: one causal, and one non-causal. I demonstrate why each approach fails to generate a satisfactory explanation, then I propose alternative non-causal possibilities and explore the viability of each. I conclude that existing approaches to both causal and non-causal explanation fail to fit the needs of XAI, though of the two approaches, non-causal accounts hold greater promise.

1.1 Deep Neural Networks

'Machine learning,'[1] an increasingly common form of AI, is a broad term that describes programs that can work with unexpected input data without being explicitly programmed to do so. One of the more common contemporary approaches to machine learning is the neural network. Neural networks attempt to replicate the behavior of biological brains by linking input and output together via various intermediary nodes in

[1] for a more comprehensive overview, see Buckner (2019).

a network. Each node is called a ‘neuron’, hence ‘neural network’. Neural networks contain multiple layers including an input layer, an output layer, and one or more ‘hidden layers’ between the input and output. Each layer is made up of a group of neurons. Neural networks with more than three hidden layers are called deep neural networks (DNNs). DNNs produce a complex, often non-interpretable model that is used in decision or classification tasks. In what is called ‘supervised learning,’ a ‘trained model’ is created by providing labeled datasets to the DNN, which iterates over the labeled data and builds a model capable of making the correct decision or classification given novel data. In other words, the deep neural model is built with the deep neural network. DNNs and the models they produce are both in need of explanation.

2 The Current Landscape: Two Case Studies

Computer scientists have made use of two contrasting strategies in order to develop XAI. Most researchers attempting to build explainable DNNs appear to prefer causal forms of explanation,[2] however some have attempted to develop non-causally explainable DNNs. I present instances of each approach and discuss their relationships to the explanation literature in the philosophy of science.

2.1 Case Study One: “Rationalizations”

One approach to XAI is to develop algorithms that produce patterns of explananda that imitate human reasoning. This is analogous to chatbots that imitate human texting

[2] See for example Yang et al. (2016), Jain and Wallace (2019), Khademi and Honavarand (2019), and Sharma, Henderson, and Ghosh (2020)

patterns. For instance, Harrison et al. (2017) uses two AIs. The first plays the classic video game Frogger, and the second explains the actions of the first by translating internal game state data to natural-language approximations of human-supplied explanations. In order to accomplish this, the research team recorded human subjects playing Frogger, then periodically paused the game and asked the subjects to verbally explain an action that they recently took. The human responses were used as training data for the "explainer" DNN.

Importantly, the explainer DNN was not generating veridical statements about the internal state of the game-playing DNN, but was generating a unique natural-language statement based on data gathered from human players when in similar in-game situations. This approach generates psychologically satisfying explanations of AI behavior. Because the generated explanations are only meant to approximate human-supplied explanations of similar situations, a tradeoff is made between accurately reporting internal DNN states and psychologically satisfying explanations. The authors accept this tradeoff in order to obtain quickly-generated and human-like explanations. The authors write that "rationalization is fast, sacrificing absolute accuracy for real-time response" (Harrison et al. 2017, 1).

The explainer DNN does not supply a veridical explanation of the decision making process used by the game-player DNN. Instead it produces statements that approximate human-generated explanations when faced with similar in-game circumstances. Another much deeper problem with this model is that, since the explanation of one DNN is itself generated by a different, independent DNN, there is now a need for an explanation of the explanation. If one black-box system is explained by appealing to a second black-box system, nothing has actually been explained. The number of phenomena in need of explanation has actually increased.

If humans depend on the use of AI for a critical task, it is important that a sense of trust in that AI is maintained. One goal of the research of Harrison et al. (2017) is to provide explanations that reassure human operators of AI that the AI had a good reason for doing an action that may appear to a human to be questionable. In some cases this may mean that the AI only needs to be able to communicate that a good reason for a particular action exists, i.e. to articulate a how-possibly explanation, rather than communicating the right reason for the action, i.e. a how-actually explanation.

Rationalizations are an attempt to deal with the problems associated with the lack of XAI without actually solving them. The authors endorse the view that, when it comes to AI, we must choose between fast, intuitive, human-understandable explanations, and technically correct explanations. Rationalizations do not attempt to provide explanations, but instead provide fictional statements that sound like plausible explanations.

2.2 Why Rationalizations are not Explanations

Rationalizations represent only one attempt to build non-causal XAI, but this attempt leaves much to be desired from the standpoint of scientific explanation. Rationalizations are explicitly non-veridical. Fictionalizations often serve a role in scientific explanation. Many, including Potochnik (2017) and Rice (2018), have argued that fictionalizations can play a key role in understanding. Rationalizations differ from fictionalizations in other models. If the understanding that an explanation helps to foster is not in any sense an understanding of a true state of affairs, then the purported explanation has not contributed to epistemic success, and is not actually explanatory. Rationalizations do not make use of strategic inaccuracies in order to help individuals to come to recognize a greater truth about the explanandum, rather rationalizations serve to further conceal the truth behind natural language statements meant to have the appearance of an adequate explanation with none of its substance. While there may be practical reasons why AI developers would find it appropriate to make use of rationalizations rather than genuine explanations, this does not imply that rationalizations have any value as scientific explanations. Rationalizations are an attempt to articulate “how possibly” explanations rather than “how actually” explanations. In the case of explanations of high-stakes automated decisions, “how actually” should be the standard. Rationalizations are not explanations.

2.3 Case Study Two: Attention Layers in Neural Networks

Attention mechanisms, introduced by Bahdanau et al. (2015), allow the training of a DNN in such a way as to focus the network’s attention on specific input elements. Attention mechanisms can be incorporated into neural networks as another layer of the network as shown in figure 1. The weights of the attention layer are thought to correlate to measures of feature importance in the input: the input has some features that are more important than others, and if the attention layer is able to identify which features of the input are most important, this is thought to generate explanantia by discriminating between relevant and irrelevant inputs. Allowing the DNN to focus on the more important parts of the input could increase the accuracy of the output. In the case of attention as explanation, the explanandum is the output of the DNN, and the explanans involves an appeal to the attention layer, which points to specific input elements. In many cases, it appeared as if the attention layer was explanatory because it indicated which parts of the input were most important in the creation of the output. For those evaluating these systems for explanatory value, this appears to be a plausible explanation, though as I will discuss, there are good reasons for doubting that this is true.

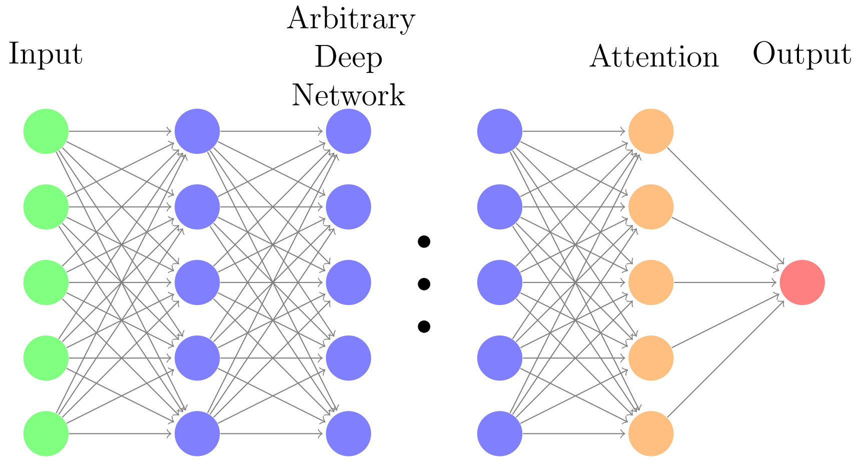

Figure 1: Researchers often use attention weights (shown in orange) to generate explanations. Jain & Wallace scramble attention weights and show that output remains stable; a similar result is obtained by Serrano & Smith omitting highly-weighted attention nodes entirely.

2.3.1 Critical Responses from Computer Science

Jain and Wallace (2019) argue that the output of the attention layer cannot serve as an explanation of the underlying DNN because it is possible to intentionally interefere with the way the weights of the attention layer are set (called "adversarial weighting") in such a way that the underlying DNN produces the same output as it did under non-adversarial weighting while the adversarial attention layer indicates the importance of entirely different - and obviously unimportant - elements of the input data. An example discussed by Jain and Wallace is the use of a DNN to gauge whether a movie review is positive or negative. The DNN outputs a number between 0 and 1 with 0 being very negative and 1 being very positive. The attention layer indicates which words in the movie review (the input) are supposedly more important in determining this output. Under the non-adversarial case, a word like "waste" would be indicated as important, whereas under the adversarial weighting, a word like "was" would be indicated as important. In both the adversarial and non-adversarial cases, the network produced an identical score for the review.

While the attention weights were set adversarially, they still represent a configuration that could have occurred during the non-adversarial training of the network. In developing a neural model under normal conditions, the production of either of the models (adversarial or non-adversarial) are equally possible. If one expects that the attention layer can serve as an explantion of the overall model, it must be the result of the ability of the attention layer to identify the most important features of the input data, but if selectively randomized attention weightings can produce the same model output as the actual attention weights, it is difficult to see in what sense the attention layer could possibly generate an explanation. Jain and Wallace (2019) conclude that it cannot. Their paper is appropriately titled "Attention is not Explanation."

Serrano and Smith (2019) make a similar argument, agreeing that attention is not explanation. Instead of assigning randomized weights to the attention nodes, Serrano and Smith selectively deleted many of the highest weighted - that is the supposedly most important - attention nodes. Under these conditions the model still produced the same output. The experiment demonstrates that if adversarial attention weightings using data that should adversely affect the neural model's accuracy has no such effect, the ability of the attention layer to discriminate between important and unimportant inputs is called into question, and so must be any explanations that are derived from attention.

Both of these papers relied on counterfactual analyses of the attention layer in order to come to thier conclusions: if the attention weights had been different in such and such a way, the attention layer would have identified a different set of input features, while the model's output would have remained unchanged. Implicitly, both are appealing to an interventionist account of explanation. They are attempting to determine the pattern of counterfactual dependence among the variables in the DNN. As I show below, due to the complexity and lack of interpretability of the systems this analysis is being applied to, the use of the interventionist account here is inappropriate, and is not likely to lead to the development of XAI.

2.4 Why Attention is not Explanation

Alisa Bokulich (2018) defines 'causal imperialism' as the view that "all scientific explanations are causal explanations" (141). There appears to be a large amount of causal imperialism in XAI - most attempts at XAI make use of causal explanations exclusively, assuming that anything other than a causal explanation is a fictionalization akin to the rationalizations described in section 2.1. Indeed, the bar for explanation under these conditions is so high that some authors have advocated for abandoning the project of developing explainable models entirely, opting instead only for models that are interpretable (Rudin 2019). There are simpler models that exist that are interpretable, such as decision trees, but they are generally less effective than more complex black box models. The tradeoff with these models is that a causal explanation can be more readily derived when a model is interpretable, because a pattern of counterfactual dependence within the model is easier to discover.

Given their complexity, a causal account of explanation that successfully explains DNNs is likely to be impossible because a pattern of counterfactual dependence cannot be located. The extremely high number of nodes in a DNN, each with an associated weight, is not human parsable. A complete account of causal relationships among nodes will also be non-parsable by humans. AI that is non-interpretable will necessarily also be non-explainable under causal accounts, because to say that a system is non-interpretable is to say that a pattern of counterfactual dependence cannot be established for that system. This follows directly from the definition of non-interpretability. A non-interpretable system is a black box system; when the inner workings of a system are unknown, the causal relationships between that system's components cannot be established. Given the failure of causal accounts in the development of XAI, non-causal accounts of explanation should be explored instead.

The criticisms of attention as explanation from Jain & Wallace and Serrano & Smith implicitly make use of an interventionst account of causal explanation similar to that proposed by Woodward (2003). Because the criticisms of attention as explanation attempt to establish the existence of empirically verifiable causal patterns that hold between the explanandum and those factors without which it would not have occurred, it fits within Woodward's framework. Woodward explains that "an intervention can be thought of as an idealized experimental manipulation which changes C 'surgically' in such a way that any change in E, should it occur, will occur only 'through' the change in C and not via some other route" (Woodward 2018, 119).

In order to determine the existence of causal relationships between variables in a system of variables, the relevant variables are subject to manipulation. Successful explanations, on this account, require that targeted manipulations of relevant system components cause changes in the output of that system when the system output is the explanandum. If manipulations of these parts cause changes to the system's output, the core elements of an explanation are already present. Because the critics of attention as explanation were able to modify seemingly relevant variables without changing the system output, they concluded that deriving an explanation from attention is inappropriate.

The criticisms of attention as explanation implicitly appealed to a view similar to the interventionist account of explanation, but one without a requirement that some variables in the system be held invariant such that the interventions on the system are surgical. Following this requirement ensures that the explanation which is eventually generated can't be superseded by another more plausible explanation related to variables which were not controlled for. In the social sciences, for example, a study of the effects of diet on longevity that does not control for income is likely to be tainted by many spurrious connections between variables that are better explained by the relationship between income and longevity than between diet and longevity. Without holding the extraneous variables invariant, the appropriate pattern of counterfactual dependence cannot be established. The absence of this requirement in the criticisms of attention as explanation may account for the results of these experiments: the discovery of nonsensical alternative explanations derived through the same means, which allowed the researchers to cast doubt on both sets of explanations. The situation does not improve significantly when surgical intervention is used; the problem with applying this approach to a DNN is that the number of interconnected nodes is so great that engaging in a surgical intervention on any one particular node is likely to be impossible as its value cannot be disentangled from the values of each other node. When making this explicit and taking this requirement into consideration, the outcome is the same - attention is not explanation - but for a different reason. In this case attention is not explanation because under the interventionist framework, it is impossible to engage in surgical intervention on a DNN, and it is thus impossible to find a pattern of counterfactual dependence among the relevant variables within the DNN.

Under the manipulability account of causal explanation, surgical intervention is a method of testing counterfactual conditionals of the form, “if I were to change X in such and such a way, the result would be Y.” Actually manipulating the value of X tests the truth of this conditional. Attention is only one part of a larger system of variables. The relevant system in this case is not attention alone, but attention in addition to the DNN itself. While both Jain and Wallace and Serrano and Smith demonstrate the possibility of engaging in surgical intervention on the attention configuration, similar interventions of the remainder of the system are not possible. When surgical intervention is impossible, all counterfactuals are rendered unintelligible since surgical intervention is in one sense merely the testing of a counterfactual conditional. To say that surgical intervention on a given system is impossible is to say that we cannot know the truth of certain counterfactual conditionals about that system.

Of the two case studies explored in section 2.1 and section 2.3, what initially appeared to be the more plausible approach (the use of causal explanations through attention mechanisms in DNNs) now appears as if it may be a dead end. While the use of rationalizations explored in section 2.1 has clear flaws, a factor motivating the approach, the desire to avoid the messy business of attempting to build causal explanations of DNNs, may have been correct. In the following section I will explore the possibility of applying non-causal explanations to DNNs.

3 Applying Non-Causal Accounts of Explanation to XAI

Both the causal and rationalization approaches to XAI have so far failed to yield good explanations of the decision process happening inside DNNs. The use of rationalizations was an attempt to build psychologically satisfying rather than veridical explanations. The attention example did appear to come closer to an acceptable conclusion. Even if the conclusion was that attention is not explanatory, the discovery of this fact advances the discussion and sets up the possibility for the discovery of other causal explanations in the future. For reasons I discuss below, the use of non-causal explanations is more appropriate for XAI.

The counterfactual theory of explanation (CTE) has causal and non-causal variants. Computer scientists have previously used causal CTE in attempts to build XAI. See, for instance, Wachter et al. (2017) and Sharma et al. (2020) These approaches suffer from many of the same problems identified by computer scientists as discussed in section 2.3.1 and by philosophers as discussed in section 2.4. Alexander Reutlinger (2018) proposes a pluralist extension of the CTE which would allow for both causal and non-causal explanations under the CTE. If it is possible to use a non-causal variant of the CTE to explain DNNs, it might be possible to overcome the objections described in sections 2.3.1

and 2.4.

Mathematical explanation, another candidate category of non-causal explanation of AI, comes, according to Colyvan et al. (2018), in two varieties: intra-mathematical and extra-mathematical. Intra-mathematical explanation is “the explanation of one mathematical fact in terms of other mathematical facts,” while extra-mathematical explanation is “the explanation of some physical phenomenon via appeal to mathematical facts” (Colyvan et al. 2018, 232). Extra-mathematical explanation holds great promise for XAI because all DNNs are mathematical. One possible problem is that the relationship between the math used to build AI models and the world is more complicated than, e.g. the relationship between the mathematics used for graph theory when representing the bridges in the city of Königsburg as a graph and the actual city of Königsburg. If an AI classifier is putting images in categories, it can be described and explained in mathematical terms, but the relevant question we seem to want answered isn't about the math, but about the connection between the math and the world. The question of how an AI knows the difference between strawberries and bananas isn't a question limited to its internal mathematical operations because it is also appealing - even if implicitly - to the actual difference between strawberries and bananas. The Seven Bridges of Königsburg problem can be solved with graph theory, but the explanation is still recognizable as representing the actual city of Königsburg. The connection between mathematics and the world in this case is clear, but it is not clear in the case of extra-mathematical explanations of AI.

The potential for the use of models as explanations has been disccused by Bokulich (2011), Batterman & Rice (2014), Morrison (2015), and Potochnik (2017) among others. Model explanations are an exciting possibility for DNNs because DNNs produce models which are used in decision and classification tasks. If models can serve as explanations, the explanation for deep DNNs could be found in the models they produce (referred to as deep neural models). One major problem with this approach is that with the types of explanatory models discussed in the philosophy of science literature, the model and the phenomena being modeled are different, but in the case of DNNs, the model is the phenomenon that needs to be explained. It is clear from the literature how a model could be explanatory of some external phenomenon, but it is not clear how a model could explain itself. It may be the case that the deep neural model explains the DNN rather than explaining itself, but then the problem of how to explain the model still remains. An explanation of the network that does not also explain the model (which is ultimately responsible for decision and classification tasks) is not enough. It isn't just the DNN which requires an explanation, but the DNN and the model it produces.

4 Conclusion

Because of the high stakes of AI-based decision and classification tasks, explanations of DNNs, deep neural models, and the decisions and classifications they produce are necessary. Computer scientists have attempted to develop explanations of these systems, but their efforts are inadequately grounded in theories of explanation. The study of scientific explanation by the philosophy of science is well suited to this task. non-causal accounts appear to have greater potential to explain DNNs than causal accounts. Non-causal variants of the CTE, extra-mathematical explanations, and model explanations all have potential to provide explanations of DNNs in the future, though more work needs to be done before this is possible. The persistent problems surrounding explanations of DNNs point to problems with existing accounts of scientific explanation and indicate the necessity for the extension of existing accounts of scientific explanation or the development of new accounts.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate.” In *Proceedings of ICLR*. 2015.

Batterman, Robert W., and Collin C. Rice. “Minimal Model Explanations.” *Philosophy of Science* 81, no. 3 (2014): 349–376. 10.1086/676677.

Bennett, Cynthia L, and Os Keyes. “What is the Point of Fairness? Disability, AI and The Complexity of Justice.” In *Workshop on AI Fairness for People with Disabilities at ACM SIGACCESS Conference on Computers and Accessibility*. 2019.

Bokulich, Alisa. “Searching for Non-Causal Explanations in a Sea of Causes.” Edited by Alexander Reutlinger and Juha Saatsi. Chap. 7 in *Explanation Beyond Causation*, 141–163. Oxford: Oxford University Press, 2018.

Buckner, Cameron. “Deep Learning: A Philosophical Introduction.” *Philosophy Compass* 14, no. 10 (2019). 10.1111/phc3.12625.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. “Predictably unequal? the effects of machine learning on credit markets.” *The Effects of Machine Learning on Credit Markets (November 6, 2018)*, 2018.

Harrison, Brent, Upol Ehsan, and Mark O. Riedl. “Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations.” *CoRR* abs/1702.07826 (2017).

Jain, Sarthak, and Byron C Wallace. “Attention is not Explanation.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 3543–3556. 2019.

Khademi, Aria, and Vasant Honavar. “Algorithmic Bias in Recidivism Prediction: A Causal Perspective.” *ArXiv* abs/1911.10640 (2019).

Mark Colyvan, John Cusbert, and Kelvin McQueen. “Two Flavours of Mathematical Explanation.” Edited by Alexander Reutlinger and Juha Saatsi. Chap. 11 in *Explanation Beyond Causation*, 231–249. Oxford: Oxford University Press, 2018.

Morrison, Margaret. *Reconstructing Reality: Models, Mathematics, and Simulations.* N.p.: Oup Usa, 2015.

Potochnik, A. *Idealization and the Aims of Science.* N.p.: University of Chicago Press, 2017.

Reutlinger, Alexander. “Extending the Counterfactual Theory of Explanation.” Edited by Alexander Reutlinger and Juha Saatsi. Chap. 4 in *Explanation Beyond Causation*, 74–95. Oxford: Oxford University Press, 2018.

Rice, Collin. “Idealized Models, Holistic Distortions, and Universality.” *Synthese* 195, no. 6 (2018): 2795–2819.

Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nature Machine Intelligence* 1, no. 5 (2019): 206–215.

Serrano, Sofia, and Noah A Smith. “Is Attention Interpretable?” In *Proceedings of ACL.* 2019.

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. "CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 166–172. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020. `10.1145/3375627.3375812`.

Wachter, Sandra, Brent D. Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *CoRR* abs/1711.00399 (2017).

Woodward, James. *Making Things Happen: A Theory of Causal Explanation.* Oxford Studies in Philosophy of Science. N.p.: Oxford University Press, 2003.

———. "Some Varieties of Non-Causal Explanation." Edited by Alexander Reutlinger and Juha Saatsi. Chap. 6 in *Explanation Beyond Causation*, 117–137. Oxford: Oxford University Press, 2018.

Yang, Diyi, Aaron Halfaker, Robert E Kraut, and Eduard H Hovy. "Who Did What: Editor Role Identification in Wikipedia." In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*, 446–455. 2016.