The Challenge of Explainable AI

Christopher Grimsley

May 4, 2019
Dept. of Philosophy
University of Kentucky

The challenge of explaining decisions made by artificial intelligence is becoming more relevant to the philosophy of science as the use of AI becomes more common. When an AI makes an important decision, those affected by the decision want an explanation, and in some cases, they are legally entitled to one. Computer scientists are aware of the importance of writing explainable AI, but the challenge has, so far, been difficult to overcome. Attempts are being made to solve the problem, but unless the scientists working on the problem are mindful of what constitutes an acceptable explanation, a situation may arise wherein an account that appears on the surface to be sufficiently explanatory fails to adequately explain. In order to ensure that explainable AI is grounded in a solid account of explanation, the philosophy of science must develop an account of explanation which applies to AI.

The problem of explainable AI highlights several issues related to explanation in the philosophy of science that are worth considering. Models of explanation since Hempel have all come under fire for various shortcomings. Though attempts have been made to rectify them, many problems remain. A survey of the history of the explanation literature in the philosophy of science is beyond the scope of this paper; readers who wish to learn more should see Hempel (1942)[1], Hempel and Oppenheim (1948)[2], Salmon (1984)[3], and Kitcher and Salmon (1989)[4].

The sustained problem of finding a suitable account of scientific explanation hints at the possibility that a single account of explanation that fits all scientific use cases may be out of reach. It also highlights concerns about the nature of explanatory force in general. If something is explanatory, it must be explanatory for *someone*. The issue of explanation in AI may do a better job than any previous example of demonstrating exactly what is wrong with the way philosophers of science have been conceptualizing explanation. Consistently, computer scientists refer to the "correct" explanation of AI, and contrast it with explanations of AI that are satisfactory to humans. By "correct," they mean roughly that, for example, some numbers served as input to a model, they were

[1]Hempel, "The Function of General Laws in History"
[2]Hempel and Oppenheim, "Studies in the Logic of Explanation"
[3]Salmon, *Scientific Explanation and the Causal Structure of the World*
[4]Kitcher and Salmon, *Scientific Explanation*

modified according to several specific mathematical formulas, and the resulting numbers were furnished by the AI as output. On a large enough scale this explanation fails to produce, even for mathematicians and computer scientists, the same sense of resolution that is produced by what are above termed "satisfactory explanations." Intuitively, the "correct" and satisfactory explanations really ought to be the same thing, but if they are not, at least one of several possibilities must be true.

First, humans have reached a level of technological sophistication wherein even specialists can no longer track the inner workings of their own technologies. This seems superficially true given that the same computer scientists who create AIs are even working on explainable AI at all. At this point the very concept of explanation becomes quite difficult to parse. At least colloquially, we often assume that if something is being explained, a process is taking place whereby a person with superior understanding of a phenomenon is describing the mechanics of a phenomenon to someone with a less well-formed understanding of the same phenomenon. This presupposes the existence of a person who possesses an understanding of the phenomenon to be explained. In cases where it is not human beings, but rather our instruments which seem, somehow, to be in possession of this understanding (if we recognize a machine generated model as a form of "understanding"), we encounter a serious problem with regard to explanation. When the only knowlege humans have of a particular phenomenon comes from a black box technology, there is no one available to explain the origin of that knowledge.

Second, there are true propositions about the world which we can access through the use of specialized equipment but which we are unable to understand currently (such that a "correct" explanation would be meaningful). It remains an open question whether this will change in the future, but it is conceivable that there exists a level of technological sophistication which would allow such a state of affairs to come into existence. What is perhaps most disturbing about the problem of explaining AI is the fact that in many cases AI produces good, empirically verifiable answers to very difficult questions which humans are unable to answer independently of the use of AI-assisted tools. This results in our ability to access true propositions about the world, yet fail to explain the nature of

those propositions. The worry is a sort of science-fiction dystopian future where machines provide answers to humans while humans lack the means to independently reach the same conclusions. In essence, this is a world where we are at the mercy of our own tools. The urgency of avoiding this possible future cannot be overstated, and this is one of the reasons why creating an account of explanation which applies to artificial intelligence is so important.

Third, the relationship between correct and satisfactory explanations was never as close as we had believed in the first place. This would appear to be an instantiation of the problem of induction - while most are fully satisfied by explanations such as "the eight ball went into the corner pocket because the cue ball hit it" this may turn out to not be the whole story. It might turn out that in the case of AI the best explanation may just be the model itself, even though out of all candidate explanations that is the least intuitive to humans. On the other hand we may need to consider the possibility that the hard distinction between correct and satisfactory explanations is the right approach precisely because science is a human activity geared towards generating explanations of phenomena *for humans*. In this case we may need to begin to have difficult conversations about the role of science in a world where AI is used not only as a tool for scientists, but is allowed to make discoveries independently of scientists.

In what follows, I will attempt to draw some conclusions about the implications that these possibilities have on the way that philosophers of science think about explanation. I will start by reviwing a recently published attempt by computer scientists to not only explain AI, but to write an AI that can explain AI. I will then trace the consequences of thinking about explanation in this way, suggesting possible features of a future theory of explanation within the philosophy of science which may be applicable to AI. I will identify three areas of concern which philosophers of science should consider when working on problems associated with scientific explanation and artificial intelligence: (1) philosophers of science must distinguish between explainability and interpretability when considering AI; (2) philosophers of science should focus on the relationship between the model and the world rather than focusing on the output of the model; (3) philosophers of science

must recognize that the domain of applicability of ML/DL/AI models is very narrow, which tends to obscure any existing explanatory value of those models.

1 Rationalizations: Natural Language Explanations of AI Behavior

Recently, when livestreams of the fire at the Notre Dame Cathedral began popping up on YouTube, an automated process was initiated by the video sharing platform: information fact-checking common 9/11 conspiracy theories was automatically posted alongside videos of the burning eight hundred year old holy site.[5] Those seeking to explain why YouTube mistook video of the fire at the medieval church in 2019 for video of the terrorist attack on the Twin Towers in 2001 might suggest that the videos did present similar imagery: both depict out-of-control flames, rising plumes of thick smoke, and strikingly, Notre Dame is marked by two towers along its west side. A coherent explanation of the error seems easy to form: the two videos are easy to confuse for one another because they contain similar features. The AI responsible for adding 9/11 fact checking to YouTube videos was simply confused by the similar features. This is an example of an explanation that is satisfactory but not correct. The AI that flagged the Notre Dame videos as 9/11 videos almost certainly does not use the concepts "out-of-control flames," "thick smoke," "two tower-like structures" at all. In fact, if this AI is like most image recognition algorithms[6], which commonly use convolutional neural nets, no concepts are involved at all. The entire process of image identification from start to finish is done on the basis of assigning values to pixels in an image, running those values through multiple hidden layers of adjustments according to various weights and biases, and landing at an output layer which identifies the image. The "correct" explanation of this error might be something like, the input values caused such and such a cluster of neurons to have activations of such and such a value, which caused the output layer to produce such and such result, which

[5] https://www.nytimes.com/2019/04/16/technology/youtube-notre-dame-fire.html

[6] I am speculating here, but because YouTube's algorithms are proprietary, everyone outside of the company not subject to an NDA is left with speculation as their only option.

was interpreted as a positive identification of a video of the terrorist attacks of 9/11. Explanations in the style of the second example above track what is actually happening within the AI while explanations in the style of the first example above are much easier for human beings to accept and digest, even though strictly speaking, they are wrong. In short, the problem is that explanations of the first sort are wrong, and explanations of the second sort do not do anything that we typically expect of good explanations.

This problem has motivated some computer scientists to attempt to develop explanations of decisions made by AIs that sound like explanations that humans are more likely to accept. One such attempt can be found in Harrison et al., "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations" (2017). The approach described in this paper is "a rationalization technique that uses neural machine translation to translate internal state-action representations of an autonomous agent into natural language."[7] In order to accomplish this, the research team recorded human subjects playing the classic video game Frogger, then periodiclaly paused the game and asked the subjects to justify an action that they recently took. These responses along with information about the internal state of the game at the time of the action in question were used as a training set for an encoder-decoder neural network. This neural network was then used to create on-the-fly rationalizations of actions in later games of Frogger. The technique used by the researchers "treats the generation of explanations as a problem of translation between ad-hoc representations of states and actions in an autonomous system's environment and natural language."[8] This is a novel approach which generates explanations of AI behavior that are satisfactory for many humans. Translating between internal states and natural language makes sense as a method of explanation, though it is clear that doing so will come at the cost of communicating the deeply technical nature of those internal states. The authors are aware of this problem, but accept it as a trade-off for quickly generated and human-like explanations. [9]

[7]Harrison et al.,"Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations," 1

[8]Harrison et al.,"Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations," 1

[9]Harrison et al.,"Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations," 1

Ultimately, what has been accomplished by Harrison et. al. is not an actual explanation of the decision making process used by the AI, but rather a second AI that outputs explanations that resemble the explanations humans would provide when faced with similar circumstances. It was not the intent of the researchers to develop rigorous scientific explanations. The explanations of the AI's choices in Frogger are similar to the common-sense explanations of the YouTube fact-checking AI. While it may be intuitive to a human that two videos look similar, or that a playable character moved left to avoid a game-ending collision, this may not actually have anything to do with the process happening within the AI's logic. Another much deeper problem with this model is that, since the explanation of the AI is itself generated by a different, independent AI, 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, and in fact the number of phenomena in need of explanation has actually increased. At the same time, the explanations produced by the explanatory AI sound reasonable to humans, and on their face appear correct even if they are not. It should be noted that the goal of the Rationalization approach to explainable AI is not to provide deep, correct, technical explanations, but to provide explanations that satisfy the human members of teams that use AI in their work. 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. 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, even if that is not technically the exact reason why the AI did what it did. It may be the case that, when it comes to AI, we must choose between intuitive, human-understandable, satisfactory explanations, and technically correct explanations. There are certainly many situations where the former is preferred over the later. One of the aims of the study of scientific explanation is to find a theory of explanation which incorporates elements both of the former and the latter - explanations which are both satisfying and correct. AI gives us good reasons to expect that such a

goal is no longer reasonable.

2 Explanation vs. Interpretation

At the start of this essay I discussed a way of thinking about explanation common to many computer scientists which says that the correct explanation of an AI decision can be found in the mathematical relationships between the parts of the model that the AI runs on. To explain an AI's action, they might say, look at the output of the model and determine how this output is related to the input via the transformations of the input as dictated by the model. The explanation of the output can be derived from the inputs to the model and the rules of the model. To a philosopher of science, this may sound a lot like the DN model. There are important reasons to think, however, that the two ways of thinking about the problem are in fact quite different.

Computer scientists distinguish between "explainability" and "interpretability." When it comes to AI, interpretability would appear to be an even bigger challenge than explanation. In some cases an explanation of an AI's actions can be constructed from the fact that the AI was working toward a particular goal and the action in question brought the AI closer to that goal. A more technical explanation may use more precise language and appeal to more math, but would essentially make the same type of claim. Interpretability is a bigger challenge because the existence of an ML model which allows an AI to perform tasks which bring it closer to the completion of a goal does not suggest anything at all about *how* the model allows the AI to achieve its desired effects.

Theories of explanation in the philosophy of science have always kept the concepts of explanation and interpretability intertwined. To the extent that an explanation under, e.g, the DN model is explanatory at all, it is both exlanatory and interpretable. Similarly, an explanation that fits the causal-mechanical model will be explanatory and interpretable in equal measures. In other words, when philosophers of science talk about explanation, they are usually talking about both explainability and interpretability at the same time. There are good reasons to do this when it comes to those areas of science like physics,

chemistry, and biology which have typically been used as examples in the explanation literature. Interpretability is not as much of an issue in these sciences. If the example being used is Newton's second law of motion, the concepts necessary to think through the example are relatively easily accessible by humans, leaving very little in the way of ambiguity in interpretation. One can easily fill in numbers for the variables in the equation and still track the things which they represent along with the relationships between them. It is not difficult to imagine the collisions of various bodies while tracking how changes to the mass or acceleration of the imaginary objects would change the outcome. If instead the example being used is a convolutional neural network with 20,000 inputs, 40 hidden layers, and a binary output, it is impossible for a human to track the relationships between all of the nodes of the network, even using sophisticated tools, making interpretation a serious problem. With the CNN example, we would need to not only explain why the output was, say, zero rather than one, but also interpret the meaning of each activation of each node of the network, the relationships between the activations at one layer and those at the next, and so on. Explanation is simple by comparison. If the concepts of explanation and interpretation have come apart in computer science, it may be helpful for philosophers of science who wish to build a new theory of explanation to distinguish between explainability and interpretability as well.

3 The Place of Models

There is a great deal of recent philosophy of science literature on scientific explanation and scientific modeling, see for example Rohwer and Rice (2016), Rice (2015), Bokulich (2011), Bailer-Jones (2003). The challenge of explainable AI could be brought closer to a resolution by thinking more about how models themselves can serve as explanations. In the philosophy of science literature on model explanations, typically the models being considered are smaller, more stable, and more easily interpretable than the models produced by many new machine learning techniques, but considering them may still prove instructive. If there are ways in which models serve as explanations, and if that applies

to AI, then we are indeed thinking about the problem of explainable AI incorrectly; since the AI is already a model it serves as its own explanation. Of course this is entirely unsatisfying and it leaves us with many unanswered questions. The only way to resolve the conflict is to think differently about what we are asking for when we ask for an explanation.

I am arguing that in most cases, when we ask for an explanation we are asking both for the means by which a conclusion was reached and for a proper interpretation of precisely how the steps taken contributed to reaching that particular conclusion rather than some other possible conclusion. When we take explanation to mean human interpretable explanation, the "correct" explanation of machine learning models does not exist in most cases. When instead we conceptualize explanation without the unstated appeal to human interpretability, those models can be explained by appealing exclusively to math. The latter approach leaves us with the much more challenging task of interpretation, but at least it resolves the explanation question.

3.1 AI as a Model

If we are to apply the philosophy of science modeling literature to AI, we must first have a proper conception of the type of model that AI is. When I refer to AI as a model, what I mean is that in many types of AI, neural networks for example, the finished product which makes the seemingly "intelligent" decisions is really a complex model that transforms input into output in the same way that a model of the solar system can produce the relative positions of the planets given certain input conditions. Say, for instance, astronomers want to know the effect that a large celestial object may have on the Moon's orbit; a model could help to determine that.

Neural networks could conceivably operate much in the same way. If researchers are trying to use automated systems to identify forested areas that are more prone to wildfire[10], they may create an AI which is capable of automatically flagging certain regions

[10]this example is meant to be hypothetical, but there are actually projects currently underway that intend to use AI to prevent wildfires. https://www.forbes.com/sites/forbestechcouncil/2018/09/24/using-technology-to-assess-wildfire-risk-and-combat-wildfires/

as areas of concern. Such a system could conceivably be trained by being fed input images from drones both of areas that did and did not suffer from wildfire in the time after the images were taken. This input data could be used to train a model that is able to distinguish the areas that did suffer from wildfire from those that did not. Images that are then fed into this model from the field as input could be idetified as either problematic or non-problematic for wildfire.

There are very important differences between these two examples, however. The model of the solar system can help to explain the relationship between the positions of the planets, laws of physics, and the input conditions. A person who wished to understand why the Moon's orbit was affected by a passing celestial object could look to the model for an explanation and expect to find the model to be a satisfactory explanation. A person who wanted to understand why the AI system flagged one area and not another as a concern for wildfire would be hard pressed to receive a satisfactory explanation from the model. The reason the AI flagged one area is that it detected similarities between the input images in the training set and the input images from the field. A more technical explanation would involve tracing the transformation of the input to the output and appealing to math: the output is such as it is as a result of the structure of the pixels on the input image and the equations that the model applies to that input. In this case nothing appears to have been explained, even though the question has been answered completely. The answer to the question of why any particular model produces the output that it does is merely that the model was constructed that way for the purpose of understanding the world, given certain inputs, a pre-determined mathematical operation takes places which produces certain outputs. This is the wrong question to ask. A better way to approach AI models is to ask how they are connected to the world: why does the AI model produce answers which appear to be correct, and is this model saying more than just something about the answer it produced, but also about the structure of the world? Philosophers of science should work towards determining if there a relationship between the AI model and the world it is answering questions about, and if so, what the nature of this relationship is. It is the model that is explanatory, not its output. An account of explanation that

applies to AI should not focus on the output of the model as much as the model itself.

3.2 AI and Model Explanations

Human scientists have labored for centuries to create models of natural phenomena that work, and in most cases, even those models that work contain idealizations and distortions. These idealizations and distortions may not matter so long as the models are effective (see Rice 2017, 2018). If this is the case, then as complicated, idealized, distorted, and uninterpretable as ML models are, the fact that they work may be enough to consider them useful in developing an understanding of the natural world. This may be an inflection point where the real problem of AI becomes most visible. If distortions and idealizations don't matter, then AI models can be explanatory because they get the right answer even though the models themselves don't obviously correspond to any natural structures in the world. At the same time, many of the answers provided by AI models are counter-intuitive, even though they may be correct, which points us back toward the model for possible clues about the process that gave rise to the answer. But the model's distortions prevent a meaningful interpretation, so what's left is a black box model that yields right answers, and which may or may not correspond to anything in the world, but is more or less impossible to interpret.

Dealing with this conundrum requires reframing our conception of models. AI models are not as generally applicable as other scientific models. In fact, they are created from a very specific set of training data. In many cases the models generalize to data from outside the training set, but in many cases this generalization does not go as far as the researchers assume it will. For example, researchers recently discovered that a common image recognition algorithm could be tricked into misidentifying images of a fire truck simply by rotating the truck slightly [11] What is important to remember when considering AI models is that the domain of applicability is extremely limited. These models can likely be applied only to a dataset that is only slightly more generalized than the data the models were trained on. If philosophers of science attempt to build an

[11] Alcorn et al., "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects", CoRR(1811.11553):2018

account of explanation that applies to AI without taking into consideration the narrowed applicability domain they will continue to struggle to develop a coherent account.

4 conclusion

For philosophers of science, explainable AI presents a serious challenge. No existing account of explanation has managed to fully satisfy questions about the relationship between AI models and the answers they seem to be able to produce. In future attempts to modify existing accounts of scientific explanation or develop new accounts, philosophers of science should be mindful of several key points with regard to artificial intelligence. First, philosophers of science must distinguish between explainability and interpretability when considering AI. Technically correct explanations of AI models appear to neatly fit the DN model, but problems remain. If we conceive of explanation only in terms of technical or mathematical explanation, we will not reach the sorts of conclusions we expect, and those that we do find will not be helpful, or in many cases, comprehensible at all. Any account of explanation must be an account of explanation *for someone*; in our case, the account of explanation we need is an account of explanation for human beings. The technical/mathematical explanations of AI may be more correct, but they have no explanatory force. The best way to deal with this challenge is to split the concept of explanation into two parts: explainability and interpretability.

Second, philosophers of science should focus on the relationship between the model and the world rather than focusing on the output of the model. AI models very frequently get the right answers to very complicated questions. These answers can be independently verified, and quite often, the degree of accuracy of the models is hard for many people to understand. The ability of AI to correctly identify in only a few seconds answers which take humans months or years represents a giant leap forward for the tools of science, but this progress comes with new challenges. Researchers have, thus far, spent a great deal of time examining the output of AI models, but in order to develop an explanation of this output we must look inside the models themselves. It may be the case that

the model which correctly answered a difficult question about the natural world has an internal structure which resembles the natural world in important ways. It may also be the case that the model contains massive distortions. We already know that the models are capable of producing correct output in either case, so the work that remains to be done lies in determining whether or not the models themselves say something true about the world.

Finally, the domain of applicability of ML/DL/AI models is very narrow, which obscures any existing explanatory value. In many cases in production use of AI, modifications to input data so small as to seem trivial to humans will break the model. This limitation sets ML/DL/AI models apart from the types of models that have been considered by philosophers of science previously. The very limited scope of the models may make a general account of explanation more difficult or impossible. Future research should take this limitation into consideration, and attempt as much as possible to limit generalizations until it is absolutely clear that they are justified.

New developments in AI have been and will continue to challenge our assumptions about the meaning of scientific explanation, and as they grow increasingly complex, may even challenge our understanding of science itself. The use of AI in science is spreading rapidly; it is incumbent upon philosophers of science to keep up with this pace if we seek to understand this new era in scientific progress.

Bibliography

Alcorn, Michael A. et al. "Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects". In: *CoRR* abs/1811.11553 (2018). arXiv: 1811.11553. URL: http://arxiv.org/abs/1811.11553.

Alvarado, Rafael and Paul Humphreys. "Big data, thick mediation, and representational opacity". In: *New Literary History* 48.4 (2017), pp. 729–749.

Bailer-Jones, Daniela M. "When scientific models represent". In: *International Studies in the Philosophy of Science* 17.1 (2003), pp. 59–74.

Bokulich, Alisa. "How scientific models can explain". In: *Synthese* 180.1 (2011), pp. 33–45.

Buckner, Cameron. "Empiricism without magic: transformational abstraction in deep convolutional neural networks". In: *Synthese* 195.12 (Dec. 2018), pp. 5339–5372. ISSN: 1573-0964. DOI: 10.1007/s11229-018-01949-1. URL: https://doi.org/10.1007/s11229-018-01949-1.

Craver, Carl F. "When Mechanistic Models Explain". In: *Synthese* 153.3 (2006), pp. 355–376.

Goldman, Alvin I. "A Causal Theory of Knowing". In: *The Journal of Philosophy* 64.12 (1967), pp. 357–372. ISSN: 0022362X. URL: http://www.jstor.org/stable/2024268.

Harrison, Brent, Upol Ehsan, and Mark O. Riedl. "Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations". In: *CoRR* abs/1702.07826 (2017). arXiv: 1702.07826. URL: http://arxiv.org/abs/1702.07826.

Hempel, Carl G. "The Function of General Laws in History". In: *The Journal of Philosophy* 39.2 (1942), pp. 35–48. ISSN: 0022362X. URL: http://www.jstor.org/stable/2017635.

Hempel, Carl G. and Paul Oppenheim. "Studies in the Logic of Explanation". In: *Philosophy of Science* 15.2 (1948), pp. 135–175.

Kitcher, P. and W.C. Salmon. *Scientific Explanation*. Minnesota studies in the philosophy of science. University of Minnesota Press, 1989. ISBN: 9781452907710. URL: https://books.google.com/books?id=5WmL6TbqCQ0C.

Rice, Collin. "Idealized Models, Holistic Distortions, and Universality". In: *Synthese* 195.6 (2018), pp. 2795–2819.

— "Models Don't Decompose That Way: A Holistic View of Idealized Models". In: *The British Journal for the Philosophy of Science* 70.1 (Aug. 2017), pp. 179–208. ISSN: 0007-0882. DOI: 10.1093/bjps/axx045. eprint: http://oup.prod.sis.lan/bjps/article-pdf/70/1/179/27987710/axx045.pdf. URL: https://doi.org/10.1093/bjps/axx045.

— "Moving Beyond Causes: Optimality Models and Scientific Explanation". In: *Nous* 49.3 (2015), pp. 589–615.

Rohwer, Yasha and Collin Rice. "How are Models and Explanations Related?" In: *Erkenntnis* 81.5 (2016), pp. 1127–1148.

Salmon, W.C. *Scientific Explanation and the Causal Structure of the World*. LPE Limited Paperback Editions. Princeton University Press, 1984. ISBN: 9780691101705. URL: https://books.google.com/books?id=2ug9DwAAQBAJ.