ANNs and Unifying Explanations: Reply to Erasmus, Brunet, and Fisher

Yunus Prasetya

Draft prior to copyediting.
Published in *Philosophy & Technology* 35(2). Please cite published version.
DOI: https://doi.org/10.1007/s13347-022-00540-4

Abstract

In a recent article, Erasmus, Brunet, and Fisher (2021) argue that Artificial Neural Networks (ANNs) are explainable. They survey four influential accounts of explanation: the Deductive-Nomological model, the Inductive-Statistical model, the Causal-Mechanical model, and the New-Mechanist model. They argue that, on each of these accounts, the features that make something an explanation is invariant with regard to the complexity of the explanans and the explanandum. Therefore, they conclude, the complexity of ANNs (and other Machine Learning models) does not make them less explainable. In this reply, it is argued that Erasmus et al. left out one influential account of explanation from their discussion: the unificationist model. It is argued that, on the unificationist model, the features that make something an explanation is sensitive to complexity. Therefore, on the unificationist model, ANNs (and other Machine Learning models) are not explainable.

It is emphasized that Erasmus et al.'s general strategy is correct. The literature on explainable Artificial Intelligence can benefit by drawing from philosophical accounts of explanation. However, philosophical accounts of explanation do not settle the problem of whether ANNs are explainable because they do not unanimously declare that explanation is invariant with regard to complexity.

In a recent article, Adrian Erasmus, Tyler Brunet, and Eyal Fisher (2021) argue that artificial neural networks (ANNs) are explainable and defend an account of interpretability. This paper provides a partial reply, specifically targeting their claim that ANNs are explainable. Erasmus et al. survey four familiar accounts of explanation from the philosophy of science—the DN model, the IS model, the CM model, and the NM model. They argue that ANNs are explainable on each of these four accounts. In response, I argue that though Erasmus et al. uses a convincing strategy to defend their thesis, they were slightly hasty in drawing their conclusion. I explore a fifth account of explanation—the unificationist model—and argue that ANNs are not explainable on this account of explanation.

1. The Indefeasibility Thesis: Explanation in Philosophy of Science

Erasmus et al. defend the Indefeasibility Thesis, defined as follows.

> *Indefeasibility Thesis*: The features that make something an explanation are invariant with respect to the complexity of both the explanans and the explanandum.

If the indefeasibility thesis is true, then there is no trade-off between accuracy and complexity on the one hand and explainability on the other. In other words, the complexity of AI systems does not make them less explainable. To defend the indefeasibility thesis, Erasmus et al. survey four influential models of explanation that have been defended in the philosophy of science literature: Hempel's *Deductive Nomological* (DN) and *Inductive Statistical* (IS) models (Hempel & Oppenheim, 1948; Hempel, 1965), Salmon's *Causal Mechanical* (CM) model (Salmon, 1984), and the *New Mechanist* (NM) models (Machamer et al., 2000; Bechtel, 2011; Craver & Darden, 2013). Erasmus et al. argue that each of the four models surveyed makes the indefeasibility thesis true.

Erasmus et al.'s strategy is laudable. They are correct to point out that the literature on ANNs and medical AI systems "often makes use of the concepts of *explainability*, *understandability* and *interpretability*, but offer little critical engagement and contain persistent disagreement on the definitions of these terms" (2021, p. 835). Using the literature on explanation from the philosophy of science helps to clarify these concepts, so that we may be more precise in discussing these topics in explainable AI. Furthermore, as far as I can tell, they are correct in their assessment that each of the four models of explanation they discuss makes the indefeasibility thesis true. That is, on the DN, IS, CM, and NM models, the complexity of the explanans does not make it any less explanatory, nor does the complexity of the explanandum make it less explainable.

However, Erasmus et al. have not discussed all of the influential models of scientific explanation in assessing the indefeasibility thesis. In particular, they did not discuss the unificationist model of explanation, defended by Michael Friedman (1974) and Philip Kitcher (1981, 1989). This is significant because, as I will argue, the unificationist model makes the

indefeasibility thesis false. For my argument, I will focus on the unificationist model of explanation as defended by Kitcher (1981, 1989).[1]

2. Unifying Explanations

Some of the most impressive explanations in the history of science unify apparently distinct phenomena: James Maxwell's theory of electromagnetism shows us that electricity and magnetism are fundamentally the same; Mendelian genetics gives us a framework for predicting the inheritance of traits in a wide range of organisms; and many more. Kitcher's (1989) unificationist model of explanation is an attempt to capture this pattern of unification. His view starts with the intuition that the most impressive explanations are those that account for a wide range of phenomena—the wider the better—using a small set of brute principles—the smaller the better. In Kitcher's words, science tells us "how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and, in demonstrating this, it teaches us how to reduce the number of types of facts we have to accept as ultimate (or brute)" (1989, p. 432).

To explain Kitcher's unificationist model, we start with the following definitions:

Schematic sentences: expressions obtained by replacing some of the nonlogical expressions occurring in a sentence with dummy letters. For instance, "Organisms homozygous for *A* develop *P*."

Filling instructions: a set of directions for replacing the dummy letters of a schematic sentence. For instance, "*A* is replaced by the name of an allele."

Schematic argument: a sequence of schematic sentences.

Classification: a set of statements describing the inferential characteristics of a schematic argument. For instance, which sentences are premises and which are conclusions.

Argument pattern: a triple consisting of a schematic argument, a set of sets of filling instructions, one for each term of the schematic argument, and a classification for the schematic argument.[2]

When the dummy letters in a schematic argument have been correctly replaced according to the filling instructions, we have a particular derivation. I shall also use terms such as *premise pattern*, *conclusion pattern*, *schematic premises*, and so on. These terms are to be understood per the definitions above.

Three factors determine how well a set of argument patterns unifies phenomena. First, a set of argument patterns unifies to the extent that they can be used to derive conclusions about a

[1] The choice to focus on Kitcher over Friedman is because Friedman's account suffers from formal difficulties, as shown in (Kitcher, 1976).

[2] These definitions and examples are taken verbatim from Kitcher (1989, p. 432).

large range of phenomena. The greater the range of phenomena derivable from the set of argument patterns, the more unifying—and therefore the more explanatory—it is.

Second, unification has to do with the number of kinds of premises needed to derive different kinds of conclusions. The most unifying argument patterns allow us to derive conclusions about a large range of phenomena from a small set of premise patterns. Additionally, the particular arguments that instantiate these patterns must be similar to one another, so that we may genuinely say that our theories allow us to derive descriptions of many phenomena from a small set of principles. This leads to the third factor: the stringency of the patterns. For an argument pattern to unify, the arguments that instantiate it must be significantly similar to one another and not merely similar in some trivial way (for instance, by exhibiting the logical form of modus ponens). This similarity is ensured by imposing restrictions on the sentences that instantiate a schematic premise. The more restrictions an argument pattern imposes on its instantiations, the more stringent it is. The more stringent an argument pattern is, the more unifying it is.

An example will help illustrate how unification works. Newtonian mechanics may be interpreted as giving us a set of argument patterns. The schematic arguments include the law of universal gravitation and the laws of motion. The filling instructions tell us to replace the dummy letters in the schematic arguments with the masses, positions, velocities, and gravitational forces of objects. This argument pattern can be used to derive descriptions of the motions of the planets in the solar system, the trajectory of projectiles near the surface of the earth, the rise of the tides, and so on.

Thus, the set of argument patterns given in Newtonian mechanics allows us to derive conclusions about a large range of phenomena. We can also note that we only need Newton's laws of motion and Newton's law of universal gravitation to derive these conclusions. This speaks to the smallness of the set of premise patterns needed to describe the phenomena. Finally, while Kitcher does not offer a precise measure of stringency, it is clear that stringency has to do with the strictness of the filling instructions of an argument pattern. For instance, Kitcher claims Aristotelian syllogisms are not stringent because the filling instructions for these argument patterns "require only that some letters be replaced with predicates, others with names" (1981, p. 518). In contrast, argument patterns from Newtonian mechanics are stringent because dummy letters can only be replaced with certain values about the positions, velocities, and masses of the relevant objects.[3] In this way, Newtonian mechanics fulfills all the criteria for being a unifying explanation.

3. The Indefeasibility Thesis and Unification

As outlined above, Erasmus et al. defend the indefeasibility thesis by arguing that it is true under the four models of explanation that they discuss. On the DN model, for instance, an explanation is a valid deductive nomological argument, with the premises being the explanans and the conclusion being the explanandum. Since neither validity nor nomicity is undermined by

[3] This requirement of stringency is designed to ensure that the particular arguments that instantiate an argument pattern are similar to one another in a substantive way, not merely in some trivial way. The more stringent an argument pattern is, the more similar the arguments that instantiate it are to one another (1981, p. 518; 1989, p. 433).

complexity, it follows that explainability, on the DN model, is not undermined by complexity. Similar arguments apply to the other three models.

Unlike the other four models, the unificationist model makes the indefeasibility thesis false. First, note that on the unificationist model, explanation comes in degrees. A set of argument patterns may be more (or less) explanatory to the extent that it unifies (or fails to unify) phenomena. This means the unificationist model makes *explanation* a vague concept. How stringent must a set of arguments be? How large can a set of schematic premises be before it fails to be unifying? Any answer we give will be arbitrary. However, there could still be obvious cases where a set of argument patterns unify phenomena and obvious cases where a set of argument patterns fail to unify and explain phenomena.[4]

Next, let us see how the unificationist model makes explanation sensitive to complexity. As outlined above, a set of argument patterns unify to the extent that (1) it can be used to derive conclusions that describe a wide range of phenomena, (2) it only requires a small number of argument patterns to derive these conclusions, (3) the argument patterns are stringent enough. The first and second conditions are sensitive to complexity. The first pertains to the complexity of the explananda. The more kinds of phenomena can be derived from an argument pattern, the more complex the explananda of that argument pattern. In this way, the complexity of the explananda contributes to explainability. The second pertains to the complexity of the explanans. The fewer kinds of premises needed to derive a range of phenomena, the less complex the explanans. In this way, the complexity of the explanans makes something less explainable.

4. Unifying Explanations and ANNs

Erasmus et al. use a medical AI system (MAIS) to illustrate their argument that ANNs are explainable. Specifically, a deep learning MAIS developed by researchers from MIT and MGH, built on a ResNet-18 CNN, and reported by Lehman et al. (2019). This MAIS is designed to evaluate the density of breast tissue based on images, and it categorizes the tissue in order to help diagnose breast cancer.[5] Erasmus et al. then argue that we can produce DN, IS, CM, and NM explanations of this MAIS, thus demonstrating that ANNs are explainable. For instance, they argue that to produce a DN explanation of the MAIS's output,

> [w]e [explain] the explanandum—here, the MAIS classifying of image *I* as classification *c*—using an explanans consisting of a law-like premises—in this case, how the weights of all relevant nodes and edges produced the output value, along with the law that an output is assigned to the most probable class—and additional information about *I*—which includes the set of input values assigned to *I*, and the output value *c* (Erasmus et al. 2021, p. 844).

Can we also provide a unifying explanation of this MAIS? To provide a unifying explanation of the MAIS, it is not enough to show that we can, in principle, supply information about the weights of the nodes and edges of the MAIS that lead to the production of output

[4] Kitcher admits that "there might be genuine indeterminacy in deciding how to weigh relative stringency, paucity of patterns and range of conclusions against one another" (1989, p. 435).
[5] In some tests, the MAIS categorizes the images in four ways: fatty, scattered, heterogenous, and dense.

values. It must also be shown that the structure of the MAIS can be interpreted as providing a unifying set of argument patterns. Specifically, the MAIS must be interpreted as giving us a small enough set of argument patterns, the argument patterns in this set must be stringent enough, and the range of conclusions derivable from this set must be wide enough.

First, let's address the range of conclusions derivable from the MAIS. ML models are typically designed for specific tasks. This MAIS is no exception. It is designed specifically to classify images of breast tissue. Lehman et al.'s (2019) MAIS is trained and tested on subjects ranging from 31 to 97 years old with no exclusions (for instance, due to prior surgery or implants). As such, it is not designed to give us outputs that cover a wide range of phenomena.

Next, we need to ascertain whether Lehman et al.'s MAIS presents us with a small enough set of stringent enough premise patterns. We can construe a single predictive process of the MAIS—from the feeding of an image as input to the production of a prediction as output—as a particular argument. The MAIS takes the pixels of an image and translates them into a large set of numbers. In this way, we can interpret the nodes of the first layer of the CNN as corresponding to dummy letters, which are replaced with numbers that represent pixels on the input image. Here, we may note that the MAIS gives us a stringent set of premise patterns, with strict filling instructions telling us that dummy letters may only be replaced by numbers that correspond to pixels in the input image.

However, the MAIS cannot be construed as giving us a small set of schematic premises. Following Erasmus et al.'s suggestion of how to interpret the MAIS as DN explainable, we construe information about the weights of all the nodes and edges and the set of input values assigned to a given image as a set of premises. The schematic premises, therefore, are simply the information about the weights of all the nodes and edges. Lehman et al. (2019) do not tell us how many nodes and edges there are in their ML model, but an ANN consisting of 17 convolution layers will have a huge number of nodes and edges. On any intuitive measure, this is a huge number of schematic premises.[6] Unless there is a smaller set of principles that can be used to derive the weights of all the nodes and edges in the MAIS, we must conclude that it does not give us a small set of premise patterns.

In short, Lehman et al.'s MAIS cannot give us unifying explanations. While it can be construed as giving us stringent argument patterns, it fails to unify because it cannot be used to derive descriptions of a wide range of phenomena, and the descriptions of phenomena that it gives us are not derived from a small set of principles. This is not an isolated case. AI systems, in general, don't use small sets of principles to produce their outputs and, therefore, do not give us unifying explanations.

5. Unification and Understanding

Erasmus et al. argue that explanation and understanding are set apart, insofar as "understanding demands satisfying some […] condition(s) which are *not dependent on the qualities of* [*an*] *explanation alone*" (2021, p. 847). Specifically, they argue that understanding depends on subjective features—possibly psychological, cognitive, or contextual. Explanation, on the other hand, does not depend on such subjective features. Thus, while explanation—on the DN, IS, CM,

[6] Of course, one can play the logical game—any large number of premises can be redescribed as a large conjunction of atomic statements. However, this kind of gerrymandering will hardly be convincing from a neutral perspective.

and NM models—is indefeasible, understanding is not. This leads them to conclude that ANNs are explainable but not understandable.

On the unificationist model, things are not so simple. To be sure, explainability is still invariant with regard to certain kinds of complexity. We may claim that the Schrödinger equation is more complex than Newton's law of universal gravitation. After all, the Schrödinger equation requires knowledge of partial differentials to understand, whereas Newton's law of universal gravitation only requires knowledge of basic mathematical operators to understand. Explanation, on the unificationist model, is invariant with regard to this kind of purely subjective complexity. However, the unificationist insists that there is an objective kind of understanding achieved through the efficient systematization of beliefs. This is the kind of understanding that the unificationist model aims to capture.[7]

6. Conclusion

Erasmus et al.'s defense of the indefeasibility thesis is not as strong as it may initially seem to be. As I've attempted to show, there is at least one influential account of explanation that makes the indefeasibility thesis false. Of course, Erasmus et al. can help themselves to some of the objections that have been raised against the unificationist model to argue that it is an inadequate model of explanation.[8] However, as Erasmus et al. note, this is also true of the four accounts of explanation they discuss.

To end, we must highlight Erasmus et al.'s contribution to the discussion of explainable AI. Erasmus et al. are correct to claim that writers who discuss AI often use concepts such as *explanation* without offering an account of explanation. This becomes problematic when these writers disagree on whether AI systems are explainable. Without an account of explanation, we are left with clashing intuitions on whether AI systems are explainable. Erasmus et al. show us how the discussion on explainable AI can benefit by using notions of explanation from the philosophy of science. However, philosophical accounts of explanation do not settle the problem of whether ANNs are explainable because they do not unanimously declare that explanation is indefeasible. Thus, I propose that Erasmus et al.'s contribution be summarized as follows. The question, "are AI systems explainable?" is too ambiguous. It is better to ask, "in what sense are AI systems explainable?" Erasmus et al. have shown that AI systems are explainable in the DN, IS, CM, and NM senses of the term. I have attempted to show that AI systems are not explainable in the unificationist sense of the term.

[7] For discussion pertaining to objective vs. subjective understanding, see (H. W. de Regt, 2009, 2013; Grimm, 2010; Lipton, 2009; Strevens, 2013).

[8] See, for examples, (Gijsbers, 2007; Shaffer, 2020).

References

Bechtel, W. (2011). Mechanism and Biological Explanation. *Philosophy of Science*, *78*(4), 533–557. https://doi.org/10.1086/661513

Craver, C. F., & Darden, L. (2013). *In Search of Mechanisms: Discoveries across the Life Sciences*. The University of Chicago Press.

de Regt, H. W. (2009). The Epistemic Value of Understanding. *Philosophy of Science*, *76*, 585–597. https://doi.org/10.1086/605795

de Regt, H. W. (2013). Understanding and Explanation: Living Apart Together? *Studies in History and Philosophy of Science*, *44*, 505–509. https://doi.org/10.1016/j.shpsa.2012.12.002

Erasmus, A., Brunet, T. D. P., & Fisher, E. (2021). What is Interpretability? *Philosophy and Technology*, *34*, 833–865. https://doi.org/10.1007/s13347-020-00435-2

Friedman, M. (1974). Explanation and Scientific Understanding. *The Journal of Philosophy*, *71*(1), 5–19. https://doi.org/10.2307/2024924

Gijsbers, V. (2007). Why Unification is Neither Necessary nor Sufficient for Explanation. *Philosophy of Science*, *74*(4), 481–500. https://doi.org/10.1086/524420

Grimm, S. R. (2010). The Goal of Explanation. *Studies in History and Philosophy of Science*, *41*, 337–344. https://doi.org/10.1016/j.shpsa.2010.10.006

Hempel, C. G. (1965). Aspects of Scientific Explanation. In *Aspects of Scientific Explanation and Other Essays* (pp. 331–496). The Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.

Kitcher, P. (1976). Explanation, Conjunction, and Unification. *The Journal of Philosophy*, *73*, 207–212. https://doi.org/10.2307/2025559

Kitcher, P. (1981). Explanatory Unification. *Philosophy of Science*, *48*(4), 507–531. https://doi.org/10.1086/289019

Kitcher, P. (1989). Explanatory Unification and the Causal Structure of the World. In P. Kitcher & W. C. Salmon, *Scientific Explanation* (pp. 410–505). University of Minnesota Press.

Lehman, C. D., Yala, A., Schuster, T., Dontchos, B., Bahl, M., Swanson, K., & Barzilay, R. (2019). Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology*, *290*(1), 52–58. https://doi.org/10.1148/radiol.2018180694

Lipton, P. (2009). Understanding Without Explanation. In H. de Regt, S. Leonelli, & K. Eigner, *Scientific Understanding: Philosophical Perspectives* (pp. 43–63). University of Pittsburgh Press.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/doi.org/10.1086/392759

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

Shaffer, M. J. (2020). Unification and the Myth of Purely Reductive Understanding. *Organon F*, *27*(2), 142–168. https://doi.org/10.31577/orgf.2020.27201

Strevens, M. (2013). No Understanding Without Explanation. *Studies in History and Philosophy of Science*, *44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005