Philosophy & Technology (2022) 35:42
https://doi.org/10.1007/s13347-022-00537-z

COMMENTARY

Interpretability and Unification

Adrian Erasmus[1,2] · T. D. P. Brunet[3]

Received: 14 April 2022 / Accepted: 18 April 2022

Abstract
In a recent reply to our article, “What is Interpretability?,” Prasetya argues against our position that artificial neural networks are explainable. It is claimed that our indefeasibility thesis—that adding complexity to an explanation of a phenomenon does not make the phenomenon any less explainable—is false. More precisely, Prasetya argues that unificationist explanations are defeasible to increasing complexity, and thus, we may not be able to provide such explanations of highly complex AI models. The reply highlights an important lacuna in our original paper, the omission of the unificationist account of explanation, and affords us the opportunity to respond. Here, we argue that artificial neural networks are explainable in a way that should satisfy unificationists and that interpretability methods present ways in which ML theories can achieve unification.

Keyword Interpretability · Explanation · Unification · XAI · Artificial neural networks

1 ANNs and Unifying Explanations

In a reply to our article, “What is Interpretability?,” Yunus Prasetya (2022) argues against our claim that artificial neural networks (ANNs) are explainable. More specifically, the paper argues that we have overlooked an influential account of explanation, the unificationist account (Kitcher, 1981, 1989), and that on this account, our indefeasibility thesis is false.

✉ Adrian Erasmus
aderasmus@ua.edu

T. D. P. Brunet
t.d.p.brunet@exeter.ac.uk

[1] Department of Philosophy, University of Alabama, 336 ten Hoor Hall, Tuscaloosa, AL 35487-0218, USA

[2] Institute for the Future of Knowledge, University of Johannesburg, Johannesburg, South Africa

[3] Department of Sociology, Philosophy, and Anthropology, University of Exeter, Amory Building, Rennes Drive, Exeter EX4 4RJ, UK

Published online: 24 April 2022

> Indefeasibility Thesis: Adding complexity to an explanation of a phenomenon does not make the phenomenon any less explainable.

Unificationist explanations, the argument goes, are defeasible to increasing complexity, and thus, such explanations of highly complex ANNs may not be available.

The reply to our article is clear and accurately describes our view. We agree with Prasetya's concluding summation of one of our contributions: "are AI systems explainable?" would be better replaced with "in what sense are AI systems explainable?" We believe this about every phenomenon, not just AI systems. We also agree that the omission of the unificationist account of explanation reveals an important lacuna in our paper. The exclusion of some accounts of explanation was necessary to constrain our account of interpretability, the primary objective of the paper, to a manageable size. Fortunately, the reply provides us with an opportunity to discuss the issue of unification as it relates to the indefeasibility thesis and interpretability.

Our response is two-part: (1) we argue that ANNs are explainable in a way that should be satisfying to unificationists and (2) we then claim that interpretability methods (and many other methods within the ML research community) are ways that ML theories achieve unification.

Before addressing the unificationist account directly, we must stress a point about explanation that we think is liable to lead to confusion about our indefeasibility thesis. There is a difference between being a "good" or "satisfying" or "understandable" explanation and being an explanation simpliciter. The reply rightly notes that many of the most "impressive" explanatory theories achieved a great deal of unification—Newtonian and Darwinian theories being two of Kitcher's chief examples. However, our indefeasibility thesis does not claim that adding complexity to an explanation of a phenomenon does not make that explanation any less good, satisfying, or impressive; we only claim that adding complexity does not make these into non-explanations. Unimpressive explanations are still explanations. The distinction between being an explanation and being a qualified explanation is essential to our account of interpretability and our response here.

2 Unificationist Explanations and the Indefeasibility Thesis

To begin, it's crucial to understand what constitutes an explanation under the unificationist account. Explanations are, according to Kitcher, arguments construed as derivations, sequences of statements in which each statement is specified as either a premise or entailment. A derivation is said to be acceptable relative to the current set of scientific beliefs, K, just in case it belongs to a set of derivations that unifies K. Unification is important because it articulates how science improves our understanding. Indeed, this is precisely what's behind Kitcher's intuition regarding one of the central aims of science (Kitcher, 1989, p. 432, emphasis ours):

> Science advances *our understanding* of nature by showing 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).

As Kitcher (1981, p. 519) notes, there are numerous possible sets that will unify K, but the best-unifying set of derivations is the one which achieves the best trade-off between minimizing the number of general argument patterns, from which derivations are instantiated, and maximizing the number of phenomena that can be explained using those patterns (Kitcher, 1989, p. 431). The upshot here is that maximizing understanding means employing those derivations belonging to the set of derivations, E(K), that best unifies the current set of scientific beliefs, K. Importantly, this does not mean that derivations belonging to other, less unifying sets of derivations no longer count as explanations. It just means that derivations, or explanations, from these other sets may not advance our understanding to the same degree as those from E(K).

The main thrust of the unificationist account is that unification is a property of a set of argument patterns that promote scientific understanding. What determines the degree of unification for a set like E(K) is the number of argument patterns it contains, and the number of phenomena for which derivations can be developed using those argument patterns. We grant, in line with Prasetya, that the addition of a large set of premises describing the workings of ANNs can increase the complexity of available argument patterns without purchasing a correspondingly large and diverse number of consequences. Indeed, in some cases, the addition of such information about an ANN will only allow us to explain the outputs of that specific ANN. In such a case, it may be that some less complex, more unifying basis for our explanatory store exists. In this sense, the additional complexity in the form of a description of an ANN can result in a less unified set of argument patterns. However, that does not mean that instantiations of those complex argument patterns are non-explanations.

Applying the indefeasibility thesis to derivations from explanatory stores over K further illustrates that such derivations are explanations regardless of whether they are in highly unifying sets. To begin, note that derivations, according to Kitcher (1989, p. 448) are DN explanations: “The explanatory store contains only deductive arguments. In a certain sense, all explanation is deductive.” In addition, Prasetya grants that our indefeasibility thesis applies to DN explanations.

Suppose we increase the complexity of some derivation from E(K) in the same way that we would another DN explanation (see our original paper). Under the unificationist account, this additional complexity would mean that the derivation is no longer an instantiation of the original argument pattern. Because of the increased complexity, the derivation is now an instantiation of a more complex argument pattern, one with additional schematic sentences in the schematic argument, and, consequently, additional sets of filling instructions and an expanded classification. Of course, a consequence of this is that the derivation may no longer be part of E(K) since the new argument pattern may be part of another, perhaps less unifying (containing more argument patterns that have a lower capacity to describe all of K), set of explanations Σ(K). Note however, that despite the lower unification possessed by Σ(K), any derivation from that set is still acceptable as an explanation by the unificationist’s lights, since it still belongs to a set of derivations that unify, albeit possibly to a lesser extent, K. In short, a unificationist should accept the indefeasibility of explanation, since it applies to each of the explanations appearing in unifying sets.

3 Interpretability Methods Achieve Unification

The elementary application of the indefeasibility thesis above is not the only way to consider our claim that ANNs are explainable on the unificationist account. In part, this is because the unificationist view is not an account of individual explanations, in the same sense that the DN, IS, CM, and NM accounts are—rather it is an assessment of the quality of theories. In those latter accounts, we are given conditions on the explanans, the explanandum, and how the former must be connected to the latter in order to have an explanation of a given type. Whereas in the unificationist account we are given (undoubtedly important and sophisticated) conditions on a whole set of argument patterns in order for them to contribute to scientific understanding. In short, unificationism is about whether a theory is "explanatory," while the other familiar accounts of explanation are about whether a given act/argument is an explanation. So, unification must be assessed not at the level of individual complex explanations of AI, but at the level of theories about complex AI.

Kitcher's targets include Newtonian dynamics and Darwinian evolutionary theory. Perhaps there is no Newton or Darwin of AI today. Should this mean that the arguments offered as explanations of the outputs of complex AI systems are not explanations? We think not. There is a large and growing body of theoretical and experimental work that aims precisely at the rigorous production of analyses of outputs of AI systems. Despite the complexity of those systems, many of the arguments used in their analyses are explanations, and indeed should be so even on a unificationist reading. This is because there is undoubtedly a degree of unification present in the theory and science of AI research.

Arguably, many argument patterns drawn from AI research and from ANNs have been highly unifying of the domain of phenomena we wish to explain with respect to the outputs of AI. Methods in AI research (e.g., gradient descent) can be construed as argument patterns and used in the analysis of multiple different AI systems, and the use of one argument pattern (method) for the explanation of numerous cases is just what unification demands. Moreover, research on ANNs often displays a wide range of payoffs gained from singular research achievements. For instance, the same basic architecture can be used for different tasks. The ResNet-18 architecture used in our original Breast Cancer Classifier example can be used across a variety of image classification tasks and can classify images into over 1000 categories included in ImageNet (He et al., 2016). Because of this wide range of use-cases, we can use information about (complex) ANN architectures to derive results about how ANNs classify multiple distinct phenomena, from computer vision to natural language processing.

Kitcher (1989, p. 501, emphasis ours) gestures toward the role of networks in unification:

> I think it entirely possible that a different system of representation might articulate the idea of explanatory unification by employing the "same way of thinking again and again" in quite a different—and possibly more revealing —way than the notions from logic that I draw on here. Kenneth Schaffner has suggested to me that *there is work in AI that can be deployed to provide the type of account I wish to give*, and Paul Churchland has urged on me the *advantages of connectionist approaches…*logical derivation may prove to be a ham-fisted way of developing the idea of explanatory unification. But, with a relatively developed account of a number of facets of explanation available, others may see how to streamline the machinery.

Moreover, and looking beyond explanation, interpretability methods are one of the unifying theoretical tools in AI research—they help streamline the explanatory machinery. Methods such as local, partial, and approximative interpretation provide ways of generating understanding of phenomena—the output of a medical AI system, for example—by connecting different explanatory tools within AI research. The approximation of an ANN by a decision tree, for instance, is a theoretical tool with application to many different ANNs. Likewise, partial interpretation can be unifying when used to simplify the explanans appearing in E(K). The piling of complex descriptions of ANNs into the explanans of a given explanation may not provide much unification and may worsen it—as Prasetya rightfully notes—but that is, we think, beside the point and not problematic for the ML community. The utility of interpretability methods in generating explanations (and sometimes new understanding of phenomena), however, is something that achieves at least a modicum of unification for current ML research. Perhaps, in this sense, the theory of interpretability methods itself is "explanatory" on a unificationist reading. We once again thank Prasetya for his fruitful engagement with our work and for providing us with an opportunity to discuss this connection.

Author Contribution AE and TDPB cowrote and approved the first draft of this article.

Funding Author AE declares no funds, grants, or other support that were received during the preparation of this manuscript.

Author TDPB has received support from the Leverhulme Trust.

Data Availability Not applicable.

Declarations

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

References

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, 770–778.

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*(4), 507–531.

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.

Prasetya, Y. (2022). ANNs and unifying explanations: Reply to Erasmus, Brunet, and Fisher. *Philosophy & Technology* (in press).