Please cite as: Veit, W. & Browning, H. (2023). Neural Networks, AI, and the Goals of Modeling. [Add Link]

Check www.walterveit.com for citation details once published

Neural Networks, AI, and the Goals of Modeling

Walter Veit

https://orcid.org/0000-0001-7701-8995

University of Bristol

Department of Philosophy, University of Bristol, BS6 6JL UK

wrwveit@gmail.com

https://walterveit.com/

Heather Browning

https://orcid.org/0000-0003-1554-7052

University of Southampton

Department of Philosophy, University of Southampton, SO17 1BJ UK

DrHeatherBrowning@gmail.com

https://www.heatherbrowning.net/

Abstract:

Deep neural networks have found many useful applications in recent years. Of particular interest have been those instances where their successes imitate human cognition and many consider artificial intelligences to offer a lens for understanding human intelligence. Here, we criticize the underlying conflation between the predictive and explanatory power of deep neural networks by examining the goals of modeling.

Main Text:

As is often the case with technological and computational progress, our newest and most sophisticated tools come to be seen as models for human cognition. What perhaps began with Gottfried Leibniz - who famously compared the mind to a mill - has a long philosophical, and now cognitivist, tradition. While it is natural to draw inspiration from technological progress to advance our understanding of the mind, unsurprisingly there are many staunch critics of the idea that the human mind should be seen as anything like a computer, with only a difference in substance. In their target article, Bowers et al. (2023) offer a compelling instance of this general criticism, arguing against recent attempts to

describe deep neural networks (DNNs) as the best models for understanding human vision (or any form of biological vision).

While DNNs have admittedly been extremely successful at classifying objects on the basis of photographs - indeed even exceeding human levels of performance in some domains - Bowers et al. essentially argue that they have very little explanatory power for human vision, due to having little in common with the mechanisms of biological vision. In order to improve our understanding of human vision, they instead advocate focusing more on explaining actual psychological findings by offering testable hypotheses.

This argument is reminiscent of many other scientific debates, such as whether artificial neural networks constitute a good model for the human brain more generally (Saxe 2021; Schaeffer et al. 2022). It also has links to long-standing discussions in the philosophy of science on the goals of science, between those that seek successful predictions and those that seek out true explanations – a debate that is sometimes framed as instrumentalists vs. realists (see Psillos 2005). While scientists may not frame their disagreement in exactly these terms, their arguments may similarly be reflective of very different attitudes towards the methodology and theoretical assumptions of their disciplines.

Our goal here is not to argue against the view provided by Bowers et al. Indeed, we strongly agree with their general argument that the predictive power of deep neural networks is insufficient to vindicate their status as models for biological vision. Even highly theoretical work has to make contact with empirical findings to promote greater explanatory power of the models. Instead, our aim here will be to take a philosophy of science perspective to examine the goals of modeling, illuminating where the disagreements between scientists in this area originate.

Firstly, there is the concern of conflating prediction with explanation. While some early philosophers of science maintained that prediction and explanation are formally (almost) equivalent, this view was quickly challenged (Rescher 1958) and today is almost universally rejected within philosophy of science. Nevertheless, in many scientific disciplines there is still a continuous and common conflation between the predictive power of a model and its explanatory power. Thus, we should not be at all surprised that many scientists have made the jump from the striking predictive success of DNNs to the bolder claim that they are representative models of human vision. While predictive power can certainly constitute one piece of good evidence for one model having greater explanatory power than another, this relationship is not guaranteed. This is especially the case when we make extrapolations from machine learning to claims about the mechanisms behind how biological agents learn and categorize the world. As Bowers et al. point out, the current evidence does not support such a generalization and instead suggests there are more likely to be dissimilar causal mechanisms underlying the observed patterns.

Secondly, as philosophers of biology have argued for the last decades, many of the properties and abilities of biological systems can be multiply realized, i.e. they can be realized through different causal mechanisms (Sober 1999; Ross 2020). Thus, the idealizations within one model may not be adequate for its application in a different target system. Just because DNNs are the first artificial intelligences we have created that approximate human levels of success in vision (or cognition) does not mean that biological systems must be operating under the same principles. Indeed, the different origins and

constraints on developing DNNs as compared with the evolution of human vision, mean that this is even less likely to be the case.

Thirdly, the authors' emphasis on controlled experiments that help us to understand mechanisms by manipulating independent variables is an important one and one that has been a common theme in recent work in the philosophy of science (e.g. Schikore 2019). This is a very different enterprise than the search for the best predictive models and AI researchers will benefit greatly from taking note of this literature. Part of the hype about AI systems has precisely been due to the confusion between predictive power and explanatory causal understanding. Prediction can be achieved through a variety of means, many of which will not be sufficiently relevantly similar to provide a good explanation.

We wish to finish by pointing out that the inadequacy of DNNs for understanding biological vision is not at all an indictment of their usefulness for other purposes. Science operates under a plurality of models and these will inevitably have different goals (Veit 2019). It is particularly interesting that DNNs have outperformed humans in some categorization tasks, since its suggests that artificial neural networks do not have to operate in the same ways as biological vision in order to imitate or even trump its successes. Indeed, there is still an important explanatory question to answer: if DNNs could constitute a superior form of visual processing, why have biological systems evolved different ways of categorizing the world? To answer these and related questions, scientists will have to seek greater collaboration and integration with psychological and neurological research, as suggested by Bowers et al. As we thus hope to have made clear here, this debate would greatly benefit by further examining its underlying methodological and philosophical assumptions as well as engaging with the literature in philosophy of science where these issues have been discussed at length.

Conflict of interest statement:

The authors have no conflicts of interest to report.

Funding Information:

This paper is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 101018533).

References

Bowers, J.S. et. al (2023). Deep Problems with Neural Network Models of Human Vision. Behavioral and Brain Sciences.

Psillos, S. (2005). Scientific realism: How science tracks truth. Routledge.

Rescher, N. (1958). On prediction and explanation. The British Journal for the Philosophy of Science, 8(32), 281-290.

Ross, L. N. (2020). Multiple realizability from a causal perspective. Philosophy of Science, 87(4), 640-662.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question?. Nature Reviews Neuroscience, 22(1), 55-67.

Schaeffer, R., Khona, M., & Fiete, I. (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. bioRxiv, 2022-08.

Schickore, J. (2019). The structure and function of experimental control in the life sciences. Philosophy of Science, 86(2), 203-218.

Sober, E. (1999). The multiple realizability argument against reductionism. Philosophy of science, 66(4), 542-564.

Veit, W. (2019). Model Pluralism. Philosophy of the Social Sciences, 50(2), 91–114.