Neural networks, AI, and the goals of modeling

Walter Veit[a] and Heather Browning[b]

[a]Department of Philosophy, University of Bristol, Bristol, UK and [b]Department of Philosophy, University of Southampton, Southampton, UK
wrwveit@gmail.com
DrHeatherBrowning@gmail.com
https://walterveit.com/
https://www.heatherbrowning.net/

doi:10.1017/S0140525X23001681, e411

Abstract

Deep neural networks (DNNs) 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 DNNs by examining the goals of modeling.

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. 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, Nelli, & Summerfield, 2021; Schaeffer, Khona, & Fiete, 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 versus 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 toward 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 DNNs 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.

First, 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 surprised at all 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.

Second, as philosophers of biology have argued for the last several decades, many of the properties and abilities of biological systems can be multiply realized, that is, they can be realized through different causal mechanisms (Ross, 2020; Sober, 1999). 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 (AIs) 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.

Third, 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., Schickore, 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 it 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.

Financial support. This study 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 program (grant agreement number 101018533).

Competing interest. None.

References

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.

Neither hype nor gloom do DNNs justice

Felix A. Wichmann[a], Simon Kornblith[b] and Robert Geirhos[b]

[a]Neural Information Processing Group, University of Tübingen, Tübingen, Germany and [b]Google Research, Brain Team, Toronto, ON, Canada
felix.wichmann@tuebingen.de
skornblith@google.com
geirhos@google.com

doi:10.1017/S0140525X23001711, e412

Abstract

Neither the hype exemplified in some exaggerated claims about deep neural networks (DNNs), nor the gloom expressed by Bowers et al. do DNNs as models in vision science justice: DNNs rapidly evolve, and today's limitations are often tomorrow's successes. In addition, providing explanations as well as prediction and image-computability are model desiderata; one should not be favoured at the expense of the other.

We agree with Bowers et al. that some of the quoted statements at the beginning of their target article about deep neural networks (DNNs) as "best models" are exaggerated – perhaps some of them bordering on scientific hype (Intemann, 2020). However, only the authors of such exaggerated statements are to blame, not DNNs: Instead of blaming DNNs, perhaps Bowers et al. should have engaged in a critical discussion of the increasingly widespread practice of rewarding impact and boldness over carefulness and modesty that allows hyperbole to flourish in science. This is unfortunate as the target article does mention a number of valid issues with DNNs in vision science and raises a number of valid concerns. For example, we fully agree that human vision is much more than recognising photographs of objects in scenes; we also fully agree there are still a number of important behavioural differences between DNNs and humans even in terms of core object recognition (DiCarlo, Zoccolan, & Rust, 2012), that is, even when recognising photographs of objects in scenes, such as DNNs' adversarial susceptibility (target article, sect. 4.1.1) or reliance on local rather than global features (target article, sect. 4.1.3). However, we do not subscribe to the somewhat gloomy view of DNNs in vision science expressed by Bowers et al. We believe that image-computable models are essential to the future of vision science, and DNNs are currently the most promising – albeit not yet fully adequate – model class for core object recognition.

Importantly, any behavioural differences between DNNs and humans can only be a snapshot in time – true as of today. Unlike Bowers et al. we do not see any evidence that future, novel DNN architectures, training data and regimes may not be able to overcome at least some of the limitations mentioned in the target article – and Bowers et al. certainly do not provide any convincing evidence why solving such tasks is beyond DNNs *in principle, that is, forever*. In just over a decade, DNNs have come a long way from AlexNet, and we still witness tremendous progress in deep learning. Until recently, DNNs lacked robustness to image distortions; now some match or outperform humans on many of them. DNNs made very different error patterns than humans; newer models achieve at least somewhat better consistency (Geirhos et al., 2021). DNNs used to be texture-biased; now some are shape-biased similar to humans (Dehghani et al., 2023). With DNNs, today's limitations are often tomorrow's success stories.

Yes, current DNNs still fail on a large number of "psychological tasks," from (un-)crowding (Doerig, Bornet, Choung, & Herzog, 2020) to focusing on local rather than global shape (Baker, Lu, Erlikhman, & Kellman, 2018), from similarity judgements (German & Jacobs, 2020) to combinatorial judgements (Montero, Bowers, Costa, Ludwig, & Malhotra, 2022); furthermore, current DNNs lack (proper, human-like) sensitivity to Gestalt principles (Biscione & Bowers, 2023). But current DNNs in vision are typically trained to recognise static images; their failure on "psychological tasks" without (perhaps radically) different training or different optimisation objectives does not surprise us – just as we do not expect a traditional vision model of motion processing to predict lightness induction or an early spatial vision model to predict Gestalt laws, at least not without substantial modification and fitting it to suitable data. To overcome current DNNs' limitations on psychological tasks we need more DNN research inspired by vision science, not just engineering to improve models' overall accuracy – here we certainly agree again with Bowers et al.

Moreover, for many of the abovementioned psychological tasks, there simply do not exist successful traditional vision models. Why single out DNNs as failures if no successful computational model exists, at least not image-computable models? Traditional "object recognition" models only model isolated aspects of object recognition, and it is difficult to tell how well they model these aspects, since only image-computable models can actually recognise objects. Here, image-computability is far