


## The scientific value of explanation and prediction

Hause Lin<sup>a,b</sup> 

<sup>a</sup>Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA and <sup>b</sup>Hill and Levene Schools of Business, University of Regina, Regina, SK, Canada  
[hauselin@gmail.com](mailto:hauselin@gmail.com)  
<https://www.hauselin.com>

doi:10.1017/S0140525X23001735, e399

### Abstract

Deep neural network models have revived long-standing debates on the value of explanation versus prediction for advancing science. Bowers et al.'s critique will not make these models go away, but it is likely to prompt new work that seeks to reconcile explanatory and predictive models, which could change how we determine what constitutes valuable scientific knowledge.

Explanatory power and predictive accuracy are different qualities, but are they inconsistent or incompatible? Bowers et al.'s critique of deep neural network models of biological vision resurfaces age-old debates and controversial questions in the history of science (Breiman, 2001; Hempel & Oppenheim, 1948). First, must an explanatory model have predictive accuracy to be considered scientifically valuable? Similarly, must a predictive model have explanatory power to have scientific value? Second, what kinds of models are better for advancing scientific knowledge, and how should we determine the scientific value of models?

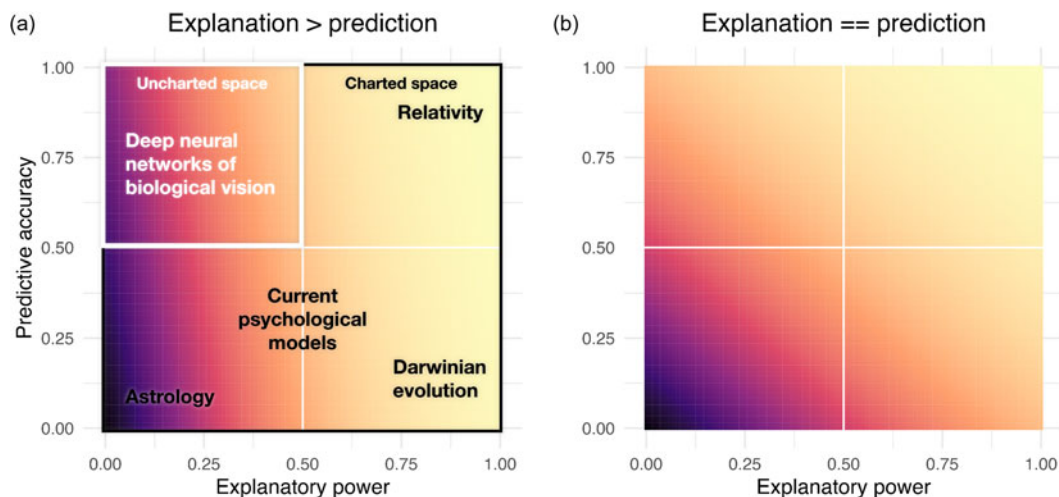
To appreciate the significance of Bowers et al.'s critique, let us consider explanation and prediction as two orthogonal dimensions rather than two extremes on a continuum. As shown in Figure 1a, some of the most successful models and theories in the history of humankind have occupied

different positions in this two-dimensional space: Theories like relativity and quantum electrodynamics are located in the top-right quadrant (i.e., very high explanatory power and predictive accuracy), whereas Darwinian evolution sits at the bottom-right quadrant (i.e., high explanatory power but little predictive accuracy, or at least cannot be tested for predictive accuracy yet). Importantly, successful models in disciplines ranging from physics to biology generally have high explanatory power.

Younger disciplines such as neuroscience and psychology – to which biological vision belongs – often aspire to emulate more established disciplines by developing models and theories with increasing explanatory power over time. Bowers et al. also prefer explanatory models and emphasize the importance of using controlled laboratory experimentation to test causal mechanisms and develop explanatory models and theories. Since researchers in these disciplines have historically valued models with explanatory power more than those with predictive accuracy, the consequence is that existing models are mostly located in the bottom two quadrants (Fig. 1a; some explanatory power but relatively low predictive accuracy). Models with high predictive accuracy are rare or even unheard of (e.g., Eisenberg et al., 2019; Yarkoni & Westfall, 2017).

Neural network models of biological vision have therefore introduced a class of scientific models that occupies a unique location in the two-dimensional space in Figure 1a (top-left quadrant). One could even argue that it might be the first time the discipline (including neuroscience and psychology) has produced models that have greater predictive accuracy than explanatory power. If so, it should come as no surprise that researchers – many of whom have been trained to rely primarily on experimentation to test theories – would feel uncomfortable with models with such different qualities and even question the scientific value of these models, despite recent calls to integrate explanation and prediction in neighboring disciplines (Hofman et al., 2021; Yarkoni & Westfall, 2017).

The current state of research on deep neural network models of biological vision reflects a critical juncture in the history of neuroscience as well as psychological and social science.



**Figure 1** (Lin). Scientific value of models with different degrees of two qualities: Explanatory power and predictive accuracy. (a) Bowers et al. value explanation over prediction, such that models with greater explanatory power are preferred. (b) Alternative value function that values both qualities equally. Hotter colors denote greater scientific value, whereas cooler colors denote less scientific value.

The long-standing tension between different philosophical approaches to theory development no longer exists only in the abstract – arguably for the first time, researchers have to reconcile, in practice, explanatory models with their predictive counterparts.

Bowers et al. emphasize the value of experimentation and the need for models to explain a wide range of experimental results. But this approach is not without limitations: When experiments and models become overly wedded to each other, models might lose touch with reality because they explain phenomena only within but not beyond the laboratory (Lin, Werner, & Inzlicht, 2021).

Should explanation be favored over prediction? The prevailing approach to theory development has certainly favored explanation (Fig. 1a), but the state of research on deep neural network models suggests that developing models with predictive accuracy might be a complementary approach that could help to test the relevance of explanatory models that have been developed through controlled experimentation. Predictive models could also be used to discover new explanations or causal mechanisms. If so, it is conceivable that current and future generations of researchers (who have been trained to also consider predictive accuracy) might come to value explanation and prediction equally (Fig. 1b).

Deep neural network models are becoming increasingly popular in a wide range of academic disciplines. Although Bowers et al.'s critique is unlikely to reverse this trend, it highlights how new methods and technological advances can turn age-old philosophical debates into practical issues researchers now have to grapple with. How the explanatory and predictive approaches are reconciled or integrated in the coming years by researchers working on biological vision is likely to have far-reaching consequences on how researchers in other disciplines think about theory development and the philosophy of science. And it is also likely to reshape our views of what constitutes valid and valuable scientific knowledge.

**Acknowledgments.** I thank Adam Bear and Alexandra Decker for helpful discussions.


**Financial support.** This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

**Competing interest.** None.

## References

- Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- Eisenberg, I. W., Bissett, P. G., Zeynep Enkavi, A., Li, J., MacKinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2019). Uncovering the structure of self-regulation through data-driven ontology discovery. *Nature Communications*, 10(1), 1–13. <https://doi.org/10.1038/s41467-019-10301-1>
- Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135–175. <https://doi.org/10.1086/286983>
- Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, 595(7866), 181–188. <https://doi.org/10.1038/s41586-021-03659-0>
- Lin, H., Werner, K. M., & Inzlicht, M. (2021). Promises and perils of experimentation: The mutual-internal-validity problem. *Perspectives on Psychological Science*, 16(4), 854–863. <https://doi.org/10.1177/1745691620974773>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>

## Fixing the problems of deep neural networks will require better training data and learning algorithms

Drew Linsley and Thomas Serre 

Department of Cognitive Linguistic & Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, RI, USA

[drew\\_linsley@brown.edu](mailto:drew_linsley@brown.edu)

[thomas\\_serre@brown.edu](mailto:thomas_serre@brown.edu)

<https://sites.brown.edu/drewlinsley>

<https://serre-lab.clps.brown.edu>

doi:10.1017/S0140525X23001589, e400

### Abstract

Bowers et al. argue that deep neural networks (DNNs) are poor models of biological vision because they often learn to rival human accuracy by relying on strategies that differ markedly from those of humans. We show that this problem is worsening as DNNs are becoming larger-scale and increasingly more accurate, and prescribe methods for building DNNs that can reliably model biological vision.

Over the past decade, vision scientists have turned to deep neural networks (DNNs) to model biological vision. The popularity of DNNs comes from their ability to achieve human-level performance on visual tasks (Geirhos et al., 2021) and the seemingly concomitant correspondence of their hidden units with biological vision (Yamins et al., 2014). Bowers et al. marshal evidence from psychology and neuroscience to argue that while DNNs and biological systems may achieve similar accuracy on visual benchmarks, they often do so by relying on qualitatively different visual features and strategies (Baker, Lu, Erlikhman, & Kellman, 2018; Malhotra, Evans, & Bowers, 2020, 2022). Based on these findings, Bowers et al. call for a reevaluation of what DNNs can tell us about biological vision and suggest dramatic adjustments going forward, potentially even moving on from DNNs altogether. Are DNNs the wrong paradigm for modeling biological vision?

### Systematically evaluating DNNs for biological vision

While this commentary identifies multiple shortcuts in DNNs that are commonly used in vision science, such as ResNet and AlexNet, it does not delve into the root causes of these issues or how widespread they are across different DNN architectures and training routines. We previously addressed these questions with *ClickMe*, a web-based game in which human participants teach DNNs how to recognize objects by highlighting category-diagnostic visual features (Linsley, Eberhardt, Sharma, Gupta, & Serre, 2017; Linsley, Shiebler, Eberhardt, & Serre, 2019). With *ClickMe*, we collected annotations of the visual features that humans rely on to recognize approximately 25% of ImageNet images (<https://serre-lab.github.io/Harmonization/>). Human feature importance maps from *ClickMe* reveal startling regularity: Animals were categorized by their faces, whereas inanimate objects like cars were categorized by their wheels and headlights