

 Access

English | Français

In response to: **Deep problems with neural network models of human vision**[Related commentaries \(29\)](#) [Author response](#)**Behavioral and Brain
Sciences****Article contents**[Abstract](#)[Financial support](#)[Competing interest](#)[References](#)

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

Published online by Cambridge University Press: **06 December 2023**In response to: **Deep problems with neural network models of human vision**[Related commentaries \(29\)](#) [Author response](#)Drew Linsley and Thomas Serre **Drew Linsley**

Affiliation: Department of Cognitive Linguistic & Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, RI,
USA drew_linsley@brown.edu thomas_serre@brown.edu
<https://sites.brown.edu/drewlinsley> <https://serre-lab.clps.brown.edu>


Thomas Serre

Affiliation: Department of Cognitive Linguistic & Psychological Sciences, Carney Institute for Brain Science, Brown University, Providence, RI,
USA drew_linsley@brown.edu thomas_serre@brown.edu
<https://sites.brown.edu/drewlinsley> <https://serre-lab.clps.brown.edu>

Commentary[Figures](#)[Related commentaries](#)[Metrics](#)**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.

Information

Type	Open Peer Commentary
Information	Behavioral and Brain Sciences, Volume 46, 2023, e400 DOI: https://doi.org/10.1017/S0140525X23001589 <div> Check for updates</div>
Copyright	Copyright © The Author(s), 2023. Published by Cambridge University Press

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 (Fig. 1a). Human participants were also significantly more accurate at rapid object classification when basing their decisions on these features rather than image saliency. In contrast, while DNNs sometimes selected the same diagnostic features as humans, they often relied on “shortcuts” for object recognition (Geirhos et al., 2020). For example, a DNN called the Vision transformer (ViT) relied on background features, like grass, to recognize a hare, whereas human participants focused almost exclusively on the animal's head (Fig. 1a). Even more concerning is that the visual features and

strategies of humans and DNNs are becoming increasingly misaligned as newer DNNs become more accurate (Fig. 1b). We and others have observed similar trade-offs between DNN accuracy on ImageNet and their ability to explain various human behavioral data and psychophysics (Fel, Felipe, Linsley, & Serre, 2022; Kumar, Houlsby, Kalchbrenner, & Cubuk, 2022). Our work indicates that the mismatch between DNN and biological vision identified by Bowers et al. is pervasive and worsening.

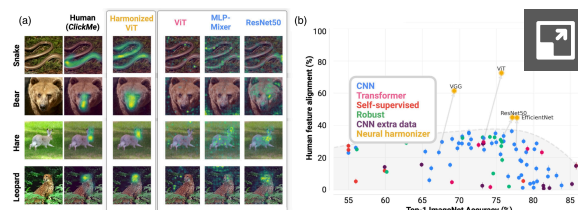


Figure 1 (Linsley and Serre). A growing misalignment between biological vision and DNNs (adapted from Fel et al., 2022). (a) Diagnostic features for object classification differ between humans and DNNs. (b) The Spearman correlation between human and DNN feature importance maps is decreasing as a function of DNN accuracy on ImageNet. This trade-off can be addressed with the neural harmonizer — a method for explicitly aligning DNN representations with humans for

object recognition.

The next generation of DNNs for biological vision

Bowers et al. argue that the inability of DNNs to learn human-like visual strategies reflects architectural limitations. They are correct that there is a rich literature demonstrating how mechanisms inspired by neuroscience can improve the capabilities of DNNs, helping them learn perceptual grouping (Kim, Linsley, Thakkar, & Serre, 2020; Linsley, Kim, Ashok, & Serre, 2019a; Linsley, Kim, Veerabadran, Windolf, & Serre, 2018, 2021), visual reasoning (Kim, Ricci, & Serre, 2018; Vaishnav et al., 2022; Vaishnav & Serre, 2023), robust object recognition (Dapello et al., 2020), and to more accurately predict neural activity (Bakhtiari, Mineault, Lillicrap, Pack, & Richards, 2021; Kubilius et al., 2018; Nayebi et al., 2018). The other fundamental difference between DNNs and biological organisms is how they learn; humans and DNNs learn from vastly different types of data with presumably different objective functions. We believe that the limitations raised by Bowers et al. result from a mismatch in data diets and objective functions because we were able to significantly improved the alignment of DNNs with humans by introducing *ClickMe* data into their training routines (“Neural harmonizer,” Fig. 1).

Biologically inspired data diets and objective functions

We believe that the power of DNNs for biological vision is from their ability to generate computational- and algorithmic-level hypotheses about vision, which will guide experiments to identify plausible circuits. For instance, the great success of gradient descent and backpropagation for training DNNs has inspired the search for biologically plausible approximations (Lillicrap, Santoro, Marris, Akerman, & Hinton, 2020). Visual neuroscience is similarly positioned to benefit from DNNs if we can improve their alignment with biology.

The most straightforward opportunity for aligning DNNs with biological vision is to train them with more biologically plausible data and objective functions (Smith & Slone, 2017; Richards et al., 2019). There have been efforts to do this with first-person video, however, these efforts have failed to yield much benefit in computer vision or other aspects of biological vision (Orhan, Gupta, & Lake, 2020; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021; Zhuang et al., 2021), potentially because the small scale of these datasets makes them ill-suited for training DNNs. An alternative approach is to utilize advances in three-dimensional (3D) computer vision, like neural radiance fields (Mildenhall et al., 2020), to generate spatiotemporal (and stereo) datasets for training DNNs that are infinitely scalable and can be integrated with other modalities, such as somatosensation and language. It is also very likely that objective functions that will lead to human-like visual strategies and features from these datasets have yet to be discovered. However, promising directions include optimizing for slow feature analysis (Wiskott & Sejnowski, 2002) and predictive coding (Lotter, Kreiman, & Cox, 2016; Mineault, Bakhtiari, Richards, & Pack, 2021), which could help align DNNs with humans without relying on *ClickMe* data.

Aligned DNNs may be all we need

Bowers et al. point out a number of ways in which DNNs fail as models of biological vision. These problems are pervasive and likely caused by the standard image datasets and training routines of DNNs, which are guided by engineering rather than biology. Bowers et al. may well be right that an entirely new class of models is needed to account for biological vision, but at the moment there are no viable alternatives. Until other model classes can rival human performance on visual tasks, we suspect that the most productive path forward toward modeling biological vision and aligning DNNs with biological vision is to develop more biologically plausible data diets and objective functions.

Acknowledgments

We thank Lakshmi Govindarajan for his helpful comments and feedback on drafts of this commentary.


Financial support


This work was supported by ONR (N00014-19-1-2029), NSF (IIS-1912280), and the ANR-3IA Artificial and Natural Intelligence Toulouse Institute (ANR-19-PI3A-0004).


Competing interest


None.


References


- 


Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, 14(12), e1006613. [CrossRef](#) [Google Scholar](#) [FindIt@Brown](#)
- 


Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34, 25164–25178. [Google Scholar](#) [FindIt@Brown](#)
- 

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 13073–13087). Curran. [Google Scholar](#) [FindIt@Brown](#)
- 


Fel, T., Felipe, I., Linsley, D., & Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., & Oh, A. (Eds.), *Advances in neural information processing systems* (Vol. 35, pp. 9432–9446). Curran Associates, Inc.
https://proceedings.neurips.cc/paper_files/paper/2022/file/3d681cc4487b97c08e5aa67224dd74f2-Paper-Conference.pdf [Google Scholar](#) [FindIt@Brown](#)
- 


Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673. [CrossRef](#) [Google Scholar](#) [FindIt@Brown](#)
- 


Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., & Wortman Vaughan, J. (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 23885–23899). Curran. [Google Scholar](#) [FindIt@Brown](#)
- 


Kim, J., Linsley, D., Thakkar, K., & Serre, T. (2020). Disentangling neural mechanisms for perceptual grouping. In Z. Chen, J. Zhang, M. Arjovsky, & L. Bottou (Eds.), *International Conference on Learning Representations*, Addis Ababa, Ethiopia. [Google Scholar](#) [FindIt@Brown](#)
- 


Kim, J., Ricci, M., & Serre, T. (2018). Not-So-CLEVR: Learning same-different relations strains feedforward neural networks. *Interface Focus*, 8(4), 20180011. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)


- 


Kubilius, J., Schrimpf, M., Nayeibi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the neural mechanisms of core object recognition. *bioRxiv*, 408385. <https://doi.org/10.1101/408385> [Google Scholar](#) [FindIt@Brown](#)
- 


Kumar, M., Houlsby, N., Kalchbrenner, N., & Cubuk, E. D. (2022). Do better ImageNet classifiers assess perceptual similarity better? [https://openreview.net > forumhttps://openreview.net > forum](https://openreview.net/forum?https://openreview.net/forum). <https://openreview.net/pdf?id=qrGKGZZvH0> [Google Scholar](#) [FindIt@Brown](#)
- 


Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews. Neuroscience*, 21(6), 335–346. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)
- 


Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., & Serre, T. (2017). What are the visual features underlying human versus machine vision? In Y. Song, C. Ma, L. Gong, J. Zhang, R. W. H. Lau, & M. Yang (Eds.), *IEEE international conference on computer vision workshops*, Venice, Italy (pp. 2706–2714). [CrossRef](#) [Google Scholar](#) [FindIt@Brown](#)
- 

Linsley, D., Kim, J., Ashok, A., & Serre, T. (2019a). Recurrent neural circuits for contour detection. *International conference on representation learning*. <https://openreview.net/forum?id=H1gB4RVKvB¬eId=H1gB4RVKvB> [Google Scholar](#) [FindIt@Brown](#)
- 

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., & Garnett, R. (Eds.), *Advances in neural information processing systems* (Vol. 31, pp. 152–164). Curran. [Google Scholar](#) [FindIt@Brown](#)
- 

Linsley, D., Malik, G., Kim, J., Govindarajan, L. N., Mingolla, E., & Serre, T. (2021). Tracking without re-recognition in humans and machines. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., & Vaughan, J. W. (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 19473–19486). Curran. [Google Scholar](#) [FindIt@Brown](#)
- 

Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2019). Learning what and where to attend. In I. Loshchilov & F. Hutter (Eds.), *7th International conference on representation learning*, New Orleans. [Google Scholar](#) [FindIt@Brown](#)
- 

Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning. *arXiv [cs.LG]*. <http://arxiv.org/abs/1605.08104> [Google Scholar](#) [FindIt@Brown](#)
- 

Malhotra, G., Dujmović, M., & Bowers, J. S. (2022). Feature blindness: A challenge for understanding and modeling visual object recognition. *PLoS*

Computational Biology, 18(5), e1009572. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Malhotra, G., Evans, B. D., & Bowers, J. S. (2020). Hiding a plane with a pixel: Examining shape-bias in CNNs and the benefit of building in biological constraints. *Vision Research*, 174, 57–68. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. *arXiv [cs.CV]*. <http://arxiv.org/abs/2003.08934> [Google Scholar](#) [FindIt@Brown](#)



Mineault, P., Bakhtiari, S., Richards, B., & Pack, C. (2021). Your head is there to move you around: Goal-driven models of the primate dorsal pathway. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., & Vaughan, J. W. (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 28757–28771). Curran. [Google Scholar](#) [FindIt@Brown](#)



Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., ... Yamins, D. L. K. (2018). Task-driven convolutional recurrent models of the visual system. *arXiv [q-bio.NC]*. <http://arxiv.org/abs/1807.00053> [Google Scholar](#) [FindIt@Brown](#)



Orhan, E., Gupta, V., & Lake, B. M. (2020). Self-supervised learning through the eyes of a child. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., & Lin, H. (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 9960–9971). Curran. [Google Scholar](#) [FindIt@Brown](#)



Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11), 1761–1770. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Smith, L. B., & Slone, L. K. (2017). A developmental approach to machine learning? *Frontiers in Psychology*, 8, 2124. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind: Discoveries in Cognitive Science*, 5, 20–29. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Vaishnav, M., Cadene, R., Alamia, A., Linsley, D., VanRullen, R., & Serre, T. (2022). Understanding the computational demands underlying visual reasoning. *Neural Computation*, 34(5), 1075–1099. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Vaishnav, M., & Serre, T. (2023). GAMR: A guided attention model for (visual) reasoning. *International conference on learning representations*.

<https://openreview.net/pdf?id=iLMgk2IGNyv> [Google Scholar](#) [FindIt@Brown](#)



Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4), 715–

770. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23), 8619–

8624. [CrossRef](#) [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)



Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences of the United States of America*, 118(3), e2014196118.

<https://doi.org/10.1073/pnas.2014196118> [Google Scholar](#) [PubMed](#) [FindIt@Brown](#)

Related content

Chapter

Teaching Computers How to See

Gabriel Kreiman

[Biological and Computer Vision](#)

Published online: 5 February 2021

Article

Deep learning: from speech recognition to language and multimodal processing

[APSIPA Transactions on Signal and Information Processing](#)

Published online: 19 January 2016

Chapter

Connectionist Models of Cognition

Michael S. C. Thomas and James L. McClelland

[The Cambridge Handbook of Computational Cognitive Sciences](#)

Published online: 21 April 2023

Article

A tutorial survey of architectures, algorithms, and applications for deep learning

Li Deng

[APSIPA Transactions on Signal and Information Processing](#)

Published online: 22 January 2014

Article

Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children†

ROMAIN SERIZEL and DIEGO GIULIANI

[Natural Language Engineering](#)

Published online: 12 April 2016

Chapter

The Modern Mathematics of Deep Learning

Julius Berner , Philipp Grohs , Gitta Kutyniok and Philipp Petersen

[Mathematical Aspects of Deep Learning](#)

Published online: 29 November 2022

Chapter

Deep Learning

Marco Gori , Frédéric Precioso and Edmondo Trentin

[The Cambridge Handbook of Computational Cognitive Sciences](#)

Published online: 21 April 2023

Chapter

Expressivity of Deep Neural Networks

Ingo Gühring , Mones Raslan and Gitta Kutyniok

[Mathematical Aspects of Deep Learning](#)

Published online: 29 November 2022

Chapter

Toward a World with Intelligent Machines That Can Interpret the Visual World

Gabriel Kreiman

[Biological and Computer Vision](#)

Published online: 5 February 2021

Chapter

Modeling Vision

Lukas Vogelsang and Pawan Sinha

[The Cambridge Handbook of Computational Cognitive Sciences](#)

Published online: 21 April 2023



Our Site

Accessibility

Contact & Help

Legal Notices

Cookie Settings

Quick Links

Cambridge Core

Cambridge Open Engage

Cambridge Aspire website

Our Products

Journals

Books

Elements

Textbooks

Join us online

Location

USA

Update