Commentary submitted to Behavioral and Brain Sciences in response to "Deep Problems with Neural Network Models of Human Vision" (Bowers et al., 2022)

Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses

Tal Golan[1], JohnMark Taylor[2], Heiko Schütt[3], Benjamin Peters[4], Rowan P. Sommers[5], Katja Seeliger[6], Adrien Doerig[7], Paul Linton[8], Talia Konkle[9], Marcel van Gerven[10], Konrad Kording[11], Blake Richards[12], Tim C. Kietzmann[13], Grace W. Lindsay[14], Nikolaus Kriegeskorte[15]

[1] Department of Cognitive and Brain Sciences, Ben-Gurion University of the Negev, golan.neuro@bgu.ac.il
[2] Zuckerman Mind Brain Behavior Institute, Columbia University, jt3295@columbia.edu
[3] Zuckerman Mind Brain Behavior Institute, Columbia University; Center for Neural Science, New York University, hs3110@columbia.edu
[4] School of Psychology & Neuroscience, University of Glasgow, benjamin.peters@posteo.de
[5] Max Planck Institute for Psycholinguistics, rowan.sommers@mpi.nl
[6] Max Planck Institute for Human Cognitive and Brain Sciences, katjaseeliger@posteo.de
[7] Institute of Cognitive Science, University of Osnabrück, adoerig@uni-osnabrueck.de
[8] Presidential Scholars in Society and Neuroscience, Italian Academy for Advanced Studies in America, and Zuckerman Mind Brain Behavior Institute, Columbia University, paul.linton@columbia.edu
[9] Department of Psychology and Center for Brain Science, Harvard University, talia_konkle@harvard.edu
[10] Donders Institute for Brain, Cognition and Behaviour, marcel.vangerven@donders.ru.nl
[11] Departments of Bioengineering and Neuroscience, University of Pennsylvania; CIFAR, koerding@gmail.com
[12] Mila; School of Computer Science, Department of Neurology & Neurosurgery, McGill University; Montreal Neurological Institute; CIFAR, blake.richards@mila.quebec
[13] Institute of Cognitive Science, University of Osnabrück, tim.kietzmann@uni-osnabrueck.de
[14] Department of Psychology and Center for Data Science, New York University, grace.lindsay@nyu.edu
[15] Zuckerman Mind Brain Behavior Institute, Department of Psychology, Department of Neuroscience, and Department of Electrical Engineering, Columbia University, nk2765@columbia.edu

Abstract: An ideal vision model accounts for behavior and neurophysiology in both naturalistic conditions and designed lab experiments. Unlike psychological theories, artificial neural networks (ANNs) actually perform visual tasks and generate testable predictions for arbitrary inputs. These advantages enable ANNs to engage the entire spectrum of the evidence. Failures of particular models drive progress in a vibrant ANN research program of human vision.

Bowers and colleagues discuss the limited connection between the psychological literature on human vision and recent work combining ANNs and benchmark-based statistical evaluation. They are correct that the psychological literature has described behavioral signatures of human vision that ANNs should but do not currently explain. A model of human vision should ideally explain all available neural and behavioral data, including the unprecedentedly rich data from naturalistic benchmarks as well as data from experiments designed to address specific psychological hypotheses. None of the current models (ANNs, handcrafted computational models, and abstractly described psychological theories) meet this challenge.

Importantly, however, the failure of current ANNs to explain all available data does not amount to a refutation of neural network models in general. Falsifying the entire, highly expressive class of ANN models is impossible. ANNs are universal approximators of dynamical systems (Funahashi & Nakamura, 1993; Schäfer & Zimmermann, 2007) and hence can implement any potential computational mechanism. Future ANNs may contain different computational mechanisms that have not yet been explored. ANNs therefore are best understood not as a monolithic falsifiable *theory* but as a *computational language* in which particular falsifiable hypotheses can be expressed. Bowers and colleagues' long list of cited studies presenting shortcomings of particular models neither demonstrates the failure of the ANN modeling framework in general nor a lack of openness of the field to falsifications of ANN models. Instead, their list of citations rather impressively illustrates the opposite: that the emerging ANN research program (referred to as "neuroconnectionism" in Doerig et al., 2022) is progressive in the sense of Lakatos: it generates a rich variety of falsifiable hypotheses (expressed in the language of ANNs) and advances through model comparison (ibid.). Each shortcoming drives improvement. For example, the discovery of texture bias in ANNs (Geirhos et al., 2019) has led to a variety of alternative training methods that make ANNs rely more

strongly on larger-scale structure in images (e.g., Geirhos et al., 2019; Hermann et al., 2020; Nuriel et al., 2020). Similarly, the discovery of adversarial susceptibility of ANNs (Szegedy et al., 2013) has motivated much research on perceptual robustness (e.g., Madry et al., 2019; Cohen et al., 2019; Guo et al., 2022).

Bowers and colleagues create a false dichotomy between benchmark studies (e.g., Schrimpf et al., 2018; Cichy et al., 2019; Nonaka et al., 2021; Kriegeskorte et al., 2008) and controlled psychological experiments. Both approaches test model-based predictions of empirical data. Traditional psychological experiments are designed to test verbally defined theories, minimizing confounders of the independent variables of theoretical interest. In contrast, the numerous experimental conditions included in natural image behavioral and neural benchmarks are high-dimensional, complex, and ecologically relevant. Controlled experiments pose specific questions. They promise to give us theoretically important bits of information but are biased by theoretical assumptions and risk missing the computational challenge of task performance under realistic conditions (Newell, 1973; Olshausen & Field, 2005). Observational studies and experiments with large numbers of natural images pose more general questions. They promise evaluation of many models with comprehensive data under more naturalistic conditions, but risk inconclusive results because they are not designed to adjudicate among alternative computational mechanisms (Rust & Movshon, 2005). Between these extremes lies a rich space of neural and behavioral empirical tests for models of vision. The community should seek models that can account for data across this spectrum, not just one end of it.

Despite their widely discussed shortcomings (e.g., Serre, 2019; Lindsay, 2021; Peters et al., 2021), ANNs are sometimes referred to as the "current best" models of human vision. This characterization is justified on both a priori and empirical grounds. A priori, ANNs are superior to verbally defined cognitive theories in that they are image-computable, i.e., they are fully computationally specified and take images as input. These properties enable ANNs to make quantitative predictions about a broad range of empirical phenomena, rendering ANNs more amenable to falsification. Being fully computationally specified enables them to make quantitative predictions of neural and behavioral responses (an advantage shared with other cognitive computational models). Taking images as inputs enables ANNs to make predictions about neural and behavioral responses to arbitrary visual stimuli. A model that explains only a particular psychological phenomenon is *a priori* inferior, ceteris paribus, to a model that predicts data across a wide range of conditions and dependent measures. The discrepancies between human vision and current ANNs are “bugs” of particular models, but the fact that we can discover these bugs is a feature of image-computable ANNs, fueling empirical progress. Since ANNs are image-computable, they enable severe tests of their predictions (superstimuli, adversarial examples, metamers; Bashivan et al., 2019; Walker et al., 2019; Dujmović, 2020; Feather et al., 2019) and powerful model comparisons (controversial stimuli; Golan et al., 2020).

The empirical reason why ANNs can be called the "current best" models of human vision is that they offer unprecedented mechanistic explanations of the human capacity to make sense of complex, naturalistic inputs. Most basically, ANNs are currently the only models that can recognize objects, parse scenes, or identify faces at performance levels similar to human performance. Furthermore, they offer image-specific predictions of errors (e.g., Rajalingham et al., 2018; Geirhos et al., 2021) and reaction times (e.g., Spoerer et al., 2017). Their predictions are far from perfect but better than those of alternative models. Finally, the intermediate representations of ANNs currently best match the neural representations that underlie human visual capacities (e.g., Güçlü & van Gerven, 2015; Dwivedi et al., 2021).

In sum, ANNs provide a language that enables us to express and test falsifiable computational models that have extraordinary power and can generalize to a broad range of empirical phenomena. Lakatos (1978) noted that all theories “are born refuted and die refuted” and stressed the importance of comparing competing theories in the light of the evidence. Our studies, then, should compare many models and report both their failures and their relative successes. It is through creation and comparison of many models that our field will progress.

References

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bowers, J., Malhotra, G., Dujmović, M., Montero, M., Tsvetkov, C., Biscione, V., . . . Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *Behavioral and Brain Sciences*, 1-74. https://doi.org/10.1017/S0140525X22002813

Cichy, R. M., Roig, G., & Oliva, A. (2019). The Algonauts Project. *Nature Machine Intelligence*, *1*(12), 613–613. https://doi.org/10.1038/s42256-019-0127-z

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified Adversarial Robustness via Randomized Smoothing. *Proceedings of the 36th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research* 97:1310–1320. https://proceedings.mlr.press/v97/cohen19c.html

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., ... & Kietzmann, T. C. (2022). The neuroconnectionist research programme. *arXiv preprint arXiv:2209.03718*. https://doi.org/10.48550/arXiv.2209.03718

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision?. *eLife*, *9*, e55978. https://doi.org/10.7554/eLife.55978

Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Computational Biology*, *17*(8), e1009267. https://doi.org/10.1371/journal.pcbi.1009267

Feather, J., Durango, A., Gonzalez, R., & McDermott, J. (2019). Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/file/ac27b77292582bc293a51055bfc994ee-Paper.pdf

Funahashi, K. I., & Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, *6*(6), 801–806. https://doi.org/10.1016/S0893-6080(05)80125-X

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. *Advances in Neural Information Processing Systems*, *34*, 23885–23899. https://proceedings.neurips.cc/paper/2021/file/c8877cff22082a16395a57e97232bb6f-Paper.pdf

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019) ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bygh9j09KX

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337. https://doi.org/10.1073/pnas.1912334117

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Guo, C., Lee, M., Leclerc, G., Dapello, J., Rao, Y., Madry, A., & Dicarlo, J.. (2022). Adversarially trained neural representations are already as robust as biological neural representations. *Proceedings of the 39th International Conference on Machine Learning*, in *Proceedings of Machine Learning Research*. https://proceedings.mlr.press/v162/guo22d.html

Hermann, K., Chen, T., & Kornblith, S. (2020). The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, *33*, 19000–19015. https://proceedings.neurips.cc/paper/2020/file/db5f9f42a7157abe65bb145000b5871a-Paper.pdf

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

Lakatos, I. (1978). Science and pseudoscience. *Philosophical Papers*, *1*, 1–7.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031. https://doi.org/10.1162/jocn_a_01544

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019) Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJzIBfZAb

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), Visual information processing: *Proceedings of the eighth annual Carnegie symposium on cognition, held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972* (pp. 283–305). Academic Press.

Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like?. *iScience*, *24*(9), 103013. https://doi.org/10.1016/j.isci.2021.103013

Nuriel, O., Benaim, S., & Wolf, L. (2021). Permuted adain: Reducing the bias towards global statistics in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9482–9491). https://openaccess.thecvf.com/content/CVPR2021/html/Nuriel_Permuted_AdaIN_Reducing_the_Bias_Towards_Global_Statistics_in_Image_CVPR_2021_paper.html

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, 5(9), 1127–1144. https://doi.org/10.1038/s41562-021-01194-6

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33), 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018

Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature Neuroscience*, 8(12), 1647–1650. https://doi.org/10.1038/nn1606

Schäfer, A. M., & Zimmermann, H. G. (2007). Recurrent neural networks are universal approximators. *International Journal of Neural Systems*, 17(04), 253–263. https://doi.org/10.1142/S0129065707001111

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like?. *BioRxiv*, 407007. https://doi.org/10.1101/407007

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, *5*, 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551. https://doi.org/10.3389/fpsyg.2017.01551

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv. arXiv:1312.6199*. https://doi.org/10.48550/arXiv.1312.6199

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., ... & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature Neuroscience*, *22*(12), 2060–2065. https://doi.org/10.1038/s41593-019-0517-x