“You can’t play 20 questions with nature and win redux”

Bradley C. Love (b.love@ucl.ac.uk, https://bradlove.org/)
University College London, United Kingdom
Robert M. Mok (Rob.Mok@mrc-cbu.cam.ac.uk,
https://sites.google.com/site/robmokbrainbob/)
University of Cambridge, United Kingdom

Abstract: 60
Main Text: 999 words
References: 152
Everything: 1285

An incomplete science begets imperfect models. Nevertheless, the target article advocates for jettisoning deep learning models with some competency in object recognition for toy models evaluated against a checklist of laboratory findings; an approach which evokes Alan Newell's 20 questions critique. We believe their approach risks incoherency and neglects the most basic test; can the model perform its intended task.

The first author remembers a discussion with fellow graduate students in the late 1990s. Each offered a prediction for when a model would be able to take photographic images as inputs and provide labels. Predictions ranged from a hundred years into the future to never. Similar estimates were provided for speech recognition. Albeit imperfect, we now have models that can perform both tasks. We marvel at the speed of progress and how poorly placed cognitive scientists were to anticipate it. In fairness, perhaps it would take that long to achieve these results if models were built by psychologists on the basis of their laboratory studies! Related discussions may have occurred in adjacent fields, such as linguistics.

In their target article, the authors correctly note some limitations of deep networks as models of vision. However, every model in an incomplete science is imperfect so these criticisms are largely benign, especially in a field that is rapidly progressing. The authors' key critique seems to be that image-computable models (i.e., models that actually attempt object recognition) are poor models of human vision because they don't account for findings from a selected set of laboratory studies. The authors invite us to return to the halcyon days before deep learning to a time of box-and-arrow models in cognitive psychology and “blocks world” models of language (Winograd, 1971), when modelers could narrowly apply toy models to toy problems safe in the knowledge that they would not be called upon to generalize beyond their confines nor pave the way for future progress.

Essentially, the authors are advocating for what Alan Newell cautioned against in his classic essay, “You can’t play 20 questions with nature and win” (Newell, 1973). Newell worried

that all the clever experiments psychologists conducted would not integrate into any coherent understanding of cognition. We agree – it seems unlikely progress will be made by amassing yet more laboratory findings. What will tie all these results together to make them more than cognitive science trivia?

One answer is models. Perhaps the most basic test for a model is whether it can perform its intended task. Once the model has some basic competency, then secondary questions can be considered, like how well the model accounts for aspects of human behavior and brain response. A model that can't pass the first hurdle, such as an object recognition model that cannot process sensory inputs (e.g., photographic images), is of little use for understanding how the brain accomplishes such feats. Models that can apply to the task can be compared on how well they account for human data (i.e., model selection). Completing the scientific loop, competing models can guide empirical efforts by suggesting informative experiments that tease apart their predictions. Instead, the authors advocate for skipping the crucial step of considering models that have basic competency and proceeding to evaluating accounts against a checklist of selected findings from laboratory studies.

This 20 questions mindset naturally pairs with the falsification approach the authors advocate. However, we do not share their enthusiasm for falsifying models that are a priori wrong and incomplete. Instead, we suggest a Bayesian or evidential philosophy of science is more appropriate in which one aims for the model that is most likely given the data (which could include data from laboratory studies). Of course, the most important empirical finding to address for a model of human object recognition is basic competency in object recognition. It seems odd to worry about fine-grain distinctions observed in the laboratory studies when the basics are missing; it is like worrying about a car's window tint when it lacks an engine and transmission.

Finally, the authors seem oddly reluctant to acknowledge or engage with work that successfully addresses their criticisms. For example, they criticize correlative approaches to assessing correspondences between brain regions and models layers, such Representation Similarity Analyses (RSA) and encoding approaches, but neglect to mention work that has successfully addressed these deficiencies. Recent work evaluates correspondences under the mantra "correlation does not imply correspondence" by directly interfacing brain activity with a model layer to evaluate whether brain activity can drive the model toward an appropriate output (i.e., behavior; Sexton & Love, 2022). Notice this approach requires a model that can perform object recognition, which further highlights the value of image-computable models in evaluating neuro-computational hypotheses. Another example is the authors' omission of large-scale "prediction" studies that successfully identify deficiencies in deep learning models and adjudicate between competing models. For example, Roads and Love (2021) derived an embedding of 50k images based on human judgments and found all deep learning models diverged from human semantic judgments with better performing models from an engineering perspective being less human aligned. This type of large-scale study provides a general and stringent test of how human aligned representations are in deep learning models. The authors mention that deep networks are susceptible to short-cut learning, which is true, but they neglect to discuss the literature devoted to ameliorating this issue, including approaches that successfully address the authors' own manipulations (e.g., adding a colored dot to an image) to create short-cuts (Dagaev et al., 2023). The

authors state that comparing models differing on a single factor is uncommon despite such comparisons being standard in machine learning papers, referred to as ablation studies. All these cases indicate that progress is being made with image-computable models on the very issues the authors highlight.

In conclusion, the fact that deep networks with some competency in object recognition fail to account for findings from some laboratory tasks has led the authors to conclude deep learning models are of limited value. One might instead conclude that the laboratory studies themselves are limited in paving the way toward a complete model of human vision. After all, our preconceived notions of how vision works guides these study designs. Some laboratory studies will prove fundamental to explaining human vision, some will be irrelevant. It seems to us that one will never be able to determine which is which in the absence of models with basic competencies.

Funding Statement

This work was supported by ESRC (ES/W007347/1), Wellcome Trust (WT106931MA), and a Royal Society Wolfson Fellowship (18302) to B.C.L., and the Medical Research Council UK (MC UU 00030/7) and a Leverhulme Trust Early Career Fellowship (Leverhulme Trust, Isaac Newton Trust: SUAI/053 G100773, SUAI/056 G105620, ECF-2019-110) to R.M.M.

Competing Interests

None.

References

Dagaev, N., Roads, B. D., Luo, X., Barry, D. N., Patil, K. R., & Love, B. C. (2023). A too-good-to-be-true prior to reduce shortcut reliance. *Pattern Recognition Letters*, 166, 164–171. https://doi.org/10.1016/j.patrec.2022.12.010

Newell, A. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. in Chase, W. G. (Ed.). (1973). Visual Information Processing: Proceedings of the Eighth Annual Carnegie Symposium on Cognition, Held at the Carnegie-Mellon University, Pittsburgh, Pennsylvania, May 19, 1972. Academic Press.

Roads, B. D., & Love, B. C. (2021). Enriching imagenet with human similarity judgments and psychological embeddings. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3547–3557.

Sexton, N. J. & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*. https://doi.org/www.science.org/doi/10.1126/sciadv.abm2219

Winograd, T. (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. *AITR-235*. http://hdl.handle.net/1721.1/7095