Disagreement and confusion over the status of DNNs as models of vision

Jeffrey S. Bowers[1], j.bowers@bristol.ac.uk; https://jeffbowers.blogs.bristol.ac.uk/

Gaurav Malhotra[1], gaurav.malhotra@bristol.ac.uk

Marin Dujmović[1], marin.dujmovic@bristol.ac.uk

Milton Llera Montero[1], m.lleramontero@bristol.ac.uk

Christian Tsvetkov[1], christian.tsvetkov@bristol.ac.uk

Valerio Biscione[1], valerio.biscione@gmail.com

Guillermo Puebla[2], guillermo.puebla@bristol.ac.uk

Federico Adolfi[3], fedeadolfi@gmail.com

John E. Hummel[4], jehummel@illinois.edu

Rachel F. Heaton[4], rmflood2@illinois.edu

Benjamin D. Evans[5], b.d.evans@sussex.ac.uk

Jeffrey Mitchell[5], j.mitchell@napier.ac.uk

Ryan Blything[6], r.blything@aston.ac.uk

[1] School of Psychological Science, University of Bristol, UK; [2] National Center for Artificial Intelligence, Vicuña Mackenna 4860, Macul, Chile; [3] Ernst Str¨ungmann Institute (ESI) for Neuroscience in Cooperation with Max Planck Society, Germany; [4] Psychology Department, University of Illinois, USA; [5] Department of Informatics, School of Engineering and Informatics, University of Sussex, UK; [6] School of Psychology, Aston University, UK.

Abstract

We are pleased there is widespread agreement that psychology has an important role to play in building better models of vision. But there are important confusions and disagreements that we discuss in our response. Our key claim is not that we should reject image-computable DNNs, but rather, image-computable DNNs should be used alongside alternative modelling frameworks to improve our understanding of vision. And when using DNNs to study human vision, a change of focus from prediction-based studies to controlled experiments that test hypotheses is needed. We also discuss a current bias in the field to focus on DNN-human similarities rather than differences. This not only contributes to a false characterization of DNN-human similarities in the domain of vision (and language and other domains), but it also impacts on research practices in ways that delays progress in building better models of mind.

Key Words: Brain-Score; Computational Neuroscience; Convolutional Neural Networks; Deep Neural Networks; Human Vision; Object identification; Object recognition; Representational Similarity Analysis.

R1. Overview

We are pleased that so many commentators agree with so many of our core claims, notably, that perceptual studies in psychology are key for evaluating DNN models of human vision; that current DNNs do a poor job in accounting for many psychological findings; that an important direction for future research is to train DNNs on new tasks and datasets that more closely capture human experience; and that new objective functions like self-supervision may improve DNN-human correspondences. Still, there are important disagreements, including the value of prediction-based studies when comparing DNNs to humans, the status of many strong claims regarding current DNN-human similarities, and the value of alternative approaches to modelling. In addition, one of the strongest objections to our work reflects a misunderstanding, namely, the view that we are advocating for rejecting DNNs altogether. We have said nothing of the sort.

The response is organized as follows. In **Section R2** we show there is no basis for the claim that we are advocating for the abandonment of DNNs as a modelling framework to test hypotheses about human vision. Rather, we are arguing that studying human vision with image computable DNNs is just one of several distinct approaches that should be pursued. In **Section R3** we consider the common claim that image computability and human-level performance on benchmark tasks are the minimal criteria for any serious model of vision. By contrast, we argue that explaining human capacities and core empirical findings are the minimum standards. In **Section R4** we consider the claim that current image computable DNNs are the "best" models of human vision. In our view, non-image computable models have contributed more to our understanding thus far and have much to offer in the future. In **Section R5** we argue that models should be developed for the sake of explanations rather than predictions. Although multiple authors highlight the importance of model predictions, there is almost no engagement with the most basic problem with prediction-based studies, namely, *Correlations do not support the conclusion that two systems share similar mechanisms (cf. Guest & Martin, 2023;* Dujmović et al., 2022*).* In **Section R6** we discuss the marketing of DNNs as the best models of human vision. In our view, the current trend of emphasizing DNN-human similarities and downplaying discrepancies is contributing to false claims and poor research practices that delay progress in building better models of human vision. In **Section R7** we specifically respond to the **DiCarlo et al.** and **Golan et al.** commentaries. Many of the (over 20) authors have played leading roles in developing this new field comparing DNNs to humans, and in both commentaries, the authors are advancing research agendas going forward. However, the authors fail to address any of our concerns, and at the same time, mischaracterize some of our key positions. Finally, in **Section R8**, we briefly summarize our arguments and conclusions.

R2. Do we recommend abandoning DNNs as models of human vision?

Many commentators claim that we are categorically rejecting DNNs as models of human vision (**Hermann et al.; Golan et al.; Love & Mok; Op De Beeck & Bracci; Summerfield & Thompson; Wichmann et al.; Yovel & Abudarham**), with quotes like:

> "In this issue of BBS, Bowers and colleagues propose that psychologists should abandon DNNs as models of human vision, because they do not produce some of the perceptual effects that are found in humans" **Yovel & Abudarham**

> "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, i.e. forever" **Wichmann et al.**

> "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" **Love & Mok**

> "…Bowers et al. take failures of ImageNet-trained models to behave in human-like ways as support for abandoning DNN architectures" **Hermann et al.**

However, this is not our position. Indeed, in Section 6.1 we clearly lay out four different approaches to modelling that should be pursued going forward, the first of which is to continue to work with standard DNNs that perform well in identifying naturalistic images of objects but modify their architectures, optimization rules, and training environments to better account for key experimental results in psychology. This is exactly the view that so many commentators are endorsing. Nowhere in the target article do we advocate for "jettisoning" DNNs, and it is hard to understand why so many researchers claim that we have.

R3. Is image computability an entry requirement for developing models of human vision?

While we explicitly endorse a research programme that, among other things, compares image computable DNNs to human vision (if severely tested), most of the commentators are less ecumenical and reject alternative modelling approaches in psychology and neuroscience that already account for some key aspects of human vision and the brain more generally. The main reason for this selective interest in DNNs is that only DNNs can recognize photographic images of objects at human or super-human levels (in some conditions), that is, only DNNs are "image computable". This is considered an essential starting point for developing models of human vision (**Anderson et al.; DiCarlo et al.; Golan et al.; Love & Mok; Op de Beeck; Spratling; Summerfield & Thompson; Wichmann et al.; Yovel & Abudarham**). As **Spratling** puts it "… the ability to process images would seem to me to be a minimum requirement for a model of vision, and models that cannot be scaled to deal with images are not worth evaluating". Similarly, **Summerfield & Thompson** describe working with non-image-computable models as "regressive". Not to be outdone, **Love & Mok** write:

> "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."

This emphasis on image computability betrays a fundamental misunderstanding of what models are and what they are for. The goal of a scientific theory/model in the cognitive sciences is to account for capacities, predict data, and explain key phenomena, not to superficially resemble that which it purports to explain. When developing DNNs of human vision, image computability makes a system *look like* a visual system, but it does not make that system a good *model* of the human visual system. The ability to identify photorealistic images is a perk, not a barrier to entry. The barrier to entry is explanatory power and accounting for key empirical results. Rather than dismiss alternative approaches to modelling because they are not image computable, the relevant questions are "what have we learned from the multitude of modelling approaches available to vision scientists?" and "what are the most promising approaches going forward?".

To answer these questions, we need to consider the different modelling approaches of the past and the different approaches currently on offer. First, there is a long history in neuroscience and psychology of developing conceptual and mathematical theories of human vision that have provided insights into key empirical phenomena, from wiring diagrams designed to explain single-cell responses of simple and complex cells in V1 (Hubel & Wiesel, 1962), to dual stream theories of vision designed to explain neuropsychological disorders of vision (Goodale & Milner, 1992), to theories of object recognition in normal vision (e.g., Biederman, 1987; Marr 1982). These approaches to modelling are still active and providing valuable insights (Baker et al., 2021; Goodale & Milner, 2023; Vannuscorps et al., 2021).

Second, there is a long history of building neural networks that process simple visual inputs to gain insights into the psychological and neural processes involved in object recognition, such as the Neocognitron model (Fukushima, 1980) that implemented and extended the theory of Hubel and Wiesel, and the JIM model that implemented and extended the theory of Biederman (Hummel & Biederman, 1992). This latter model, JIM, and its successors (Hummel & Stankiewicz, 1996; Hummel, 2001) recognize simple line drawings of objects and are premised on the assumption that the goal of the ventral visual stream is to build a representation of the distal stimulus (the world and the objects in it) that can be used to understand the visual world. On this view, object classification is merely a consequence, not the be-all and end-all, of the ventral visual stream. Unlike current DNNs, JIM, and its successors account for many key psychological findings in human object recognition--such as the sensitivity of humans to part-whole relations--without being able to process naturalistic photographic images.

In a similar way, Grossberg and colleagues developed Adaptive Resonance Theory (ART) models that quickly learn to classify simple visual patterns without forgetting past learning, that is, networks that solve the stability-plasticity dilemma (e.g., Carpenter & Grossberg, 1987; Grossberg, 1980). ART models not only account for a range of empirical findings reported in psychology and neuroscience (Grossberg, 2020), they have also been used to solve engineering challenges (Da Silva et al., 2019). Grossberg has also developed detailed models of low-level vision that take in simple visual inputs to capture a wide range of perceptual illusions (Grossberg, 2009). Expanding on the work of Grossberg, Francis et al. (2017) implemented networks that process simple visual inputs to explain a range of crowding phenomena that current DNNs cannot explain. In related work, George et al., (2017, 2020) developed Recursive Cortical Networks that support the recognition of "captchas" and can account for several phenomena core to human vision, including some Gestalt phenomena (George et al, 2018). These models rely on segmentation and occlusion-reasoning in a unified framework to support object recognition, but only work with simple visual stimuli. These modelling efforts (and many others) largely fall into the second research programme we endorse in Section 6.1, namely, building networks that focus on explaining key psychological phenomena rather than image computability.

Third, there are active research programmes today following the third approach we endorse in section 6.1, namely, building models that support various human capacities that current DNNs struggle with (without focusing on the details of psychological or neuroscience research). But again, these models cannot process the photographic images that DNNs recognize. For example, Hinton, a co-author of AlexNet, rejects current image computable DNNs as models of human vision and is instead developing Capsule and GLOM models (Hinton, 2022; Sabour et al., 2017). Hinton (2022) writes:

> There is strong psychological evidence that people parse visual scenes into part-whole hierarchies and model the viewpoint-invariant spatial relationship between a part and a whole as the coordinate transformation between intrinsic coordinate frames that they assign to the part and the whole [Hinton, 1979]. If we want to make neural networks that understand images in the same way as people do, we need to figure out how neural networks can represent part-whole hierarchies.

Indeed, current DNNs fail to represent objects in terms of their parts and relations even when explicitly trained to do so (Malhotra et al., in press).

Similarly, generative models, such as variational autoencoders, are being developed that learn disentangled representations of visual elements of a scene (single hidden units that encode shape, color, position, etc.; e.g., Higgins et al., 2016; Montero et al., 2022; Zhang et al., 2022) and object-centric learning models are being built to perform perceptual grouping (e.g., Locatello et al., 2020; Singh et al., 2021; Anciukevicius et al., 2023). To understand these principles, these models are frequently trained and tested on datasets of artificially created simple visual stimuli. **German & Jacobs** explicitly argue that variational autoencoders provide a more promising framework for understanding how human vision encodes objects in terms of their parts and relations between parts. But at present, exploring this requires working with simple rather than the photorealistic images.

The important point to emphasize here is that all these models would (and some actually do) receive very low Brain Scores (some cannot even be tested) because they cannot process the photorealistic inputs in ImageNet. Yet they constitute useful tools to explore important phenomena in constrained settings. Are we supposed to discard such approaches because they cannot process and recognize photographs of objects? We think not. In our view, the diversity of modelling approaches in psychology (and the cognitive sciences more generally) fits well with the diversity of productive questions that can be asked about cognitive systems (cf. van Rooij, 2022). This is important to counteract the assumption that all worthwhile models of vision can recognize naturalistic photographs of objects or are on a trajectory towards becoming image computable.

R4. Are image computable models the current "best" models of human vision.

Still, it might be argued that image-computable DNNs that perform well on prediction-based experiments are the current best models of human vision because they provide more insights into human vision. However, we are struggling to see what the new insights are (although see our responses to **Anderson et al**. and **Op de Beeck & Bracci** below). Current DNNs account for few findings from psychology, and only do well on brain prediction-based studies when there is no attempt to rule out confounds as the basis of their successes. At the same time, DNNs that vary in terms of their architectures (CNNs vs. Transformers), and objective functions (classification vs. image reconstruction) support similar levels of predictions on behavioural and brain benchmarks (e.g., Storrs et al. 2021), with **Hermann et al.** and **Linsey & Serre** noting a recent trend for better performing models of object recognition doing more poorly on Brain-Score (although **Wichmann et al.** note that a transformer model trained on four billion images does much better on behavioral benchmarks). And as noted by **Xu & Vaziri-Pashkam**, when RSA is assessed with higher quality brain data, the correspondence across levels of DNNs and visual cortex is lost for familiar objects, and the predictivity scores go down dramatically for unfamiliar objects. Most problematically, **Xu & Vaziri-Pashkam** note that RSA scores are greatly reduced following theoretically motivated experimental manipulations of images. What conclusions or insights about human vision follow from these observations? At present, it seems that the main advantage of image computable DNNs compared to alternative models is that they recognize things, with little evidence that they do this in the way that humans do.

In fact, many commentators readily concede that current DNNs are doing a poor job in accounting for the results of experimental studies of human vision, and multiple possible solutions have been proposed. DNNs need to be trained with a better diet of images that more closely resemble human experience (**Linsley & Serre; Op de Beeck & Bracci; Yovel & Abudarham**), more biological constraints need to be added to models, such as representing binocular input from two eyes (**Chandran et al**.), and new objective functions and tasks need to be explored, including building DNNs that support vision for action (**German & Jacobs; Hermann et al.; Li & Mur; Liu & Bartolomeo; Rothkopf et al; Slagter; Summerfield & Thomson),** with many of these authors advocating for some combination of the above approaches. Again, we agree with these research agendas, and we are pursuing some of these ourselves, including adding biological constraints to networks (Evans et al., 2022; Tsvetkov et al., 2023) and modifying training environments (Biscione & Bowers, 2022), in an attempt to make DNNs encode information in a more human-like manner. At the same time, there are good a priori reasons to think major architectural innovations may be necessary, for example, to encode relations between parts (**Kellman et al**.), with some authors pessimistically assessing the promise of DNNs as models of brains, with quotes such as: "DNN are not just inadequate models of the visual system but are so different in their structure and functionality that they are not even on the same playing field" (**Gur**) and the claim that DNNs "are doomed to be largely useless models for psychological research on language" (**Bever et al.**).

Of course, the human visual system is an image computable neural network (although a network that differs from current DNNs in many fundamental ways; Izhikevich 2004). However, the claim that current image computable DNNs are the most promising models of human vision going forward, despite the limited insights gathered thus far, is nothing more than a faith-based prophecy that may or may not pan out. In our view, researchers should be pursuing multiple different modelling approaches to advance our understanding of human vision. It is the dismissal of alternative approaches that is regressive (cf., Rich et al., 2021 for a computational account of why this is detrimental).

R5. The role of prediction and explanation in model building.

In the target article we distinguished between uncontrolled, prediction-based studies that often highlight DNN-human similarities and controlled experiments that often highlight dissimilarities. We argued that the former experiments are problematic given that predictions can be driven by confounds (e.g., a DNN that classifies images based on small-patch confounds can make good predictions on a standard neural dataset; Dujmović et al., 2023). By contrast, controlled experiments can help rule out confounds and allow researchers to draw causal conclusions regarding similarities and differences between DNNs and humans. To our surprise, few commentators address the concern that good predictions on prediction-based studies can be supported by models that transform their inputs in very different ways to humans. The exceptions are **Srivastava et al.** who highlight that similar issues apply in other domains and **Veit & Browning** who point out that properties and abilities of biological systems can be multiply realized and that controlled experiments are needed make causal conclusions regarding the similarity of DNNs and humans.

Still, multiple commentators did discuss the relation between prediction and explanation (**Lin; Moldoveanu; Op de Beeck et al.; Veit & Browning; Wichmann et al**.; **Yovel & Abudarham**), claiming that explanation and prediction serve different functions and that both are important. For example, **Wichmann et al.** write: "we believe that both prediction and explanation are required: an explanation without prediction cannot be trusted, and a prediction without explanation does not aid understanding", and **Lin** writes "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".

These comments seem to suggest that explanations derived from models do not rest on predictions. In fact, both prediction-based studies and controlled experiments test model-based predictions (**Golan et al**). The important distinction is between predictions with and without explanation. In the case of prediction-based studies, there is no manipulation of independent variables designed to test specific hypotheses regarding how DNNs make their predictions, and accordingly, no explanation for any good predictions. Indeed, receiving 100% predictivity does not help the scientist understand how a DNN is predicting (see Figure 5). By contrast, in the case of controlled experiments, where a model is assessed in how well it predicts performance across conditions designed to test hypotheses, good predictions can contribute to an explanation.

Of course, some types of predictions provide a stronger test of a model than others **(Spratling**), and this applies to both prediction-based studies and controlled experiments. In the case of prediction-based studies, current DNNs only perform well in the easy cases, namely, when training and test images are from the same distribution (often described as independent and identically distributed data or i.i.d. data). When models are assessed on their ability to make behavioural or brain predictions for test images from a different distribution (out-of-distribution data or o.o.d. data), performance plummets. For example, as noted above, **Xu &Vaziri-Pashkam** showed that brain predictivity with RSA was much weaker when they included novel stimuli in the test set. Similarly, the claim that current DNNs can perform human-like same-different visual judgements is based on testing models on images that are similar to the training set. When the images are different, performance plummets, even though the judgements are trivial for humans (Puelba & Bowers, 2022, 2023). These are all specific examples of a more general limitation of current DNNs, namely, models do poorly on out-of-distribution images. In other words, not only do prediction-based studies provide little insight into how models predict, but also their successful predictions are highly circumscribed.

Similarly, in the case of controlled experiments, successful predictions tend to fall near the space of existing results. That is, models tested on controlled experiments make the prediction that the results will replicate on another sample of participants, images, etc. taken from the same population (i.i.d. data). And just as DNNs tend to fail in their predictions on o.o.d. data, models rarely make counter-intuitive predictions that are subsequently confirmed in a controlled experiment. It is worth noting that models tested on controlled experiments are generally described as *accounting for* (rather than *predicting*) results when successful, and this terminology might be more appropriate for prediction-based studies relying on i.i.d. data.

Whatever the terminology, prediction-based studies and controlled experiments both assess how well DNNs predict (account) for data, but only the latter method tests hypotheses to rule out confounds and to make causal claims regarding how DNNs and humans identify objects.

Arguments regarding the relative advantages of prediction vs. explanation touch on a broader debate regarding the relative advantages of studying natural systems in artificial conditions that allow precise control of variables vs. naturalistic conditions where control is more limited. For example, **Love & Mok** cite the classic paper by Newell (1973) "You can't play 20 questions with nature and win" as a fundamental problem with studying the brain with controlled experiments. According to **Love & Mok**, laboratory studies in psychology have only produced a collection of findings they characterize as "cognitive science trivia". **Summerfield & Thompson** are not so dismissive of these experimental results, but they are critical of models in psychology that narrowly focus on explaining a small set of laboratory findings. DNNs, by contrast, are thought to hold promise of "genuine predictive power in the natural world" when trained on tasks that humans face in everyday life.

It strikes us as peculiar to characterize the empirical findings from psychology as "trivia" rather than core constraints for theory building and odd to dismiss models of specific empirical findings if they help explain key aspects of vision. What other area of science does not break down complex phenomena into parts? When **Summerfield & Thompson** criticize models that only have a narrow scope, writing "For example, a model that explains crowding typically does not explain filling in and vice versa", it is important to note that current DNNs account for neither.

For the sake of argument, let us accept the claim that image-computable models provide the best way forward for addressing Newell's challenge. Nevertheless, it is still the case that only controlled experiments provide specific hypotheses about how to improve DNN-human correspondences. For example, controlled experiments highlighted specific limitations of current DNNs as models of human vision (e.g., relying too much on texture, etc.) leading to specific suggestions about how to address them (e.g., a generative rather than discriminative objective function may result in a model that encodes shape rather than texture; **German & Jacobs**). A research programme of training image computable DNNs on naturalistic datasets without running specific controlled experiments will simply lead to black-box models in which there is no understanding of how the model works, let alone whether the model learns similar representations to humans.

It is also important to recognize the challenges with working with naturalistic images even when relying on controlled studies. For example, Rust and Movshon (2005) argued for the importance of building theories of biological vision using artificial and simple stimuli. They pushed back on the view that the best way to understand vision was to probe the system with naturalistic images, writing:

> "Implicit in this approach is the assumption that synthetic stimuli are in some way impoverished or 'simplistic' and therefore somehow miss important features of visual response. The main—and in our view, crippling—challenge is that the statistics of natural images are complex and poorly understood. Without understanding the constituents of natural images, it is imprudent to use them to develop a well-controlled hypothesis-driven experiment."

Although these comments were made before the current interest in DNNs, it remains just as difficult to design well-controlled hypothesis-driven experiments using natural images now as it was then given the billions of features associated with images. As a result, DNNs trained on these images become liable to learning based on short-cuts (Geirhos et al., 2020) and confounds (Dujmović, et al., 2023), making it difficult to interpret their mechanisms and internal representations.

It is also important to emphasize that model predictions are not the only way to advance our understanding of natural systems. **Lin** gives the example of Darwinian evolution as a model that has explanatory power but limited predictive accuracy. We think the term theory rather than model is more appropriate here, but the critical point is that evolution explains existing data very well, and it would be silly to dismiss the theory because it does not make precise predictions going forward. This point generalises to all areas of science, such that unimplemented theories of vision can provide important insights into human vision if they can provide an account of key existing findings. Indeed, simply running experiments that test hypotheses can be highly informative. Of course, formal modelling has an important role to play, but in all cases, the focus should be on explanation, not prediction.

R6. The marketing of DNNs as the current best models of human vision is impeding our progress in developing better models.

To advance our understanding of human vision with DNNs we need to use methods that allow us to make causal claims regarding how both DNNs and humans process visual information – that is, use controlled experiments. In addition, when running controlled experiments, it is important that researchers systematically explore both the successes and failures of DNNs to capture aspects of human vision. This involves not only correctly characterizing the results from both DNNs and humans, but also carrying out studies that attempt to falsify claims regarding DNN-human similarities. Indeed, the best empirical evidence for a model is that it survives "severe" tests (Mayo, 2018), namely, experiments that have a high probability of falsifying a claim if and only if the claim is false in some relevant manner.

However, this does not characterize standard practice in the field at present. Instead, there appears to be a bias towards highlighting similarities and downplaying differences. **Tarr** notes that many of the strong claims regarding DNN-human similarities are best understood as marketing rather than serious scientific claims – and on his view, the problem rests with the *consumers* who take the hype (too) seriously. He writes a story of a fool buying a pig because he saw a brochure suggesting pigs could fly. It is an allegory – the person should not be so naïve to believe the marketing. Similarly, he cautions us to be smart consumers of science and not take strong claims regarding DNN-human similarity too seriously. He writes that DNNs are only "proxy models" of vision and writes: "I don't think there is much actual confusion that DNNs are 'models of the human visual system'".

We imagine it would be hard for **DiCarlo et al.**, and **Golan et al.** to agree with this conclusion given they both repeat the claim that DNNs are the best models of human vision. But more importantly, this marketing impacts the field in two general ways.

R6.1. Marketing and research practices. When looking for DNN-human similarities, there is little motivation to move away from prediction-based studies that can provide misleading estimates of similarities, little reason for researchers to carry out controlled studies that provide severe tests of these claims, and little interest from editors and reviewers in publishing studies that highlight DNN-human dissimilarities. Consistent with these claims, two commentators explicitly minimize the importance of falsification. **Tarr** writes: "...less handwringing about what current models can't do; instead, they should focus on what DNNs can do". Similarly, **Love & Mok** write: "…we do not share their enthusiasm for falsifying models that are a priori wrong and incomplete". Instead, **Love & Mok** advocate for a Bayesian approach to model evaluation, where the question is which model is most likely given the data. But model selection depends on which data are under consideration, and currently, too many fundamental psychological findings are ignored because DNNs do not capture them. If Bayesian methods were used to select models that account for psychological phenomena, then in many cases, non-image computable models would perform best.

Perhaps the above comments regarding falsification are anomalous, and **Golan et al**. are right to doubt a bias in the field, but in our experience, this attitude towards falsification is widespread. For example, see the following NeurIPS workshop talk by Bowers (2022) that provides multiple examples of reviewers and editors stating that falsification is not enough. Rather, it is necessary to find "solutions" to make DNNs more like humans to publish: https://slideslive.com/38996707/researchers-comparing-dnns-to-brains-need-to-adopt-standard-methods-of-science. Similar biases are well recognized in other fields. For example, it is analogous to a bias against publishing null results in psychology that is well understood to have led to many false conclusions (Simmons et al., 2011).

R6.2. Marketing and characterizing research findings. There is another respect in which this marketing manifests itself, namely, weak or ambiguous findings are too often characterized as supporting strong conclusions. We gave multiple examples of this in the target article (e.g., Caucheteux et al., 2021; Duan et al., 2019; Kim et al., 2021; Hermann et al., 2020; Messina et al., 2021; Zhou & Firestone, 2020) and there are more examples from the current commentaries themselves. For instance, **de Vries et al.** criticize us for claiming that colour and form are processed entirely separately in V1 and cite some studies of theirs that show that DNNs do a good job in capturing important features of human colour processing. We take the point that the strong claims by Livingstone and Hubel (1988) need to be qualified given subsequent work (e.g., Garg et al., 2019), but **de Vries et al.** mischaracterize their own findings. They claim that categorical perception of colour emerges as a function of training models to classify objects and note that this effect did not emerge in a DNN trained to distinguish artificial from human-made scenes (de Vries et al. 2022). However, as reported in Appendix 7 of de Vries et al. (2022), an untrained DNN also showed some degree of categorical perceptual effects as well. This latter finding substantially weakens the evidence for their claim that colour perception emerges as a consequence of learning to classify objects.

Similarly**, Love & Mok** criticize us for not "engaging with work that successfully addresses their criticisms", but the evidence they report do not support their conclusions. **Love & Mok** give two examples from their own lab. First, they describe the work of Sexton & Love (2022) who note that RSA and linear prediction methods of comparing DNNs to brains rely on correlations and write: "Just as correlation does [not] imply causation, correlation does not imply correspondence". We agree. The problem is in how they draw correspondence claims. The authors assess whether brain signals can causally drive object recognition in DNNs by substituting the response elicited in an internal layer of a DNN with (a linear transform of) the brain response elicited by the same visual stimulus. They find that the activities from brain regions do indeed drive DNN object recognition performance above chance levels and take this as evidence that the representations in DNNs and brain are similar.

However, there are both empirical and logical problems with their studies and the conclusions they draw. Empirically, as reported in the supplemental materials (Figure S10 and Table S3), when brain data are used to drive DNN object recognition, performance drops from ~80% to < 10% in one experiment and from ~58% to < 2% in the second experiment. This large drop in performance is problematic for their conclusion. More fundamentally, the observation that brain responses support (limited) object recognition in DNNs does not address the issue of confounds. Just as texture-like representations in DNNs might be used to predict shape representations in cortex (leading to good RSA or Brain-Scores in the absence of similar representations), it is possible that shape representations in cortex can be mapped to texture-like representations in DNNs to drive object recognition to a limited extent. That is, the (weak) causal link between brain activation and DNN object recognition does nothing to address our concern that good predictions do not imply similar representations. Just as correlations do not imply causation, causation does not imply correspondence.

Love & Mok also describe a study by Dagaev et al. (2023) that they claim addresses a problem identified by Malhotra et al. (2021), namely, that DNNs are so susceptible to short-cut learning that they will classify the images from CIFAR10 based on a single pixel confound. Their solution involved introducing a *too-good-to-be-true* prior during training—if an image could be classified successfully by a low-capacity network (which Dagaev et al. use as a short-cut detector), the image is down-weighted during training a full-capacity network. This way, the full-capacity network only learned on images that, Dagaev et al. claim, are less likely to contain short-cuts. While this method is certainly of interest for a machine learning engineer, it is of limited relevance to a cognitive scientist and does not address the criticisms made by Malhotra et al. (2021). Firstly, if the shortcut is widely prevalent in the dataset—in Malhotra et al. a diagnostic pixel was present in 80-100% of images—this method would fail. Secondly, there is nothing to say that short-cuts picked up by DNNs are necessarily easier to pick up by a low-capacity network. There could be many complex short-cuts, involving a conjunction of features that will be ignored by humans and picked up by full-capacity DNNs, but not by low-capacity DNNs. The point that Dagaev et al. miss is that we do not want models to ignore simple diagnostic visual features (humans rely on heuristics across a wide range of domains) but that they should learn *the right kind of* features i.e., models should incorporate appropriate human inductive biases, not whatever the low-capacity DNN does not happen to find diagnostic.

Yovel & Abudarham describe how DNNs capture the face-inversion effect, writing: “Interestingly, a human-like face inversion effect that is larger than an object inversion effect is found in DNNs”. In fact, as shown by Yovel et al. (2022) and others, DNNs show similar size inversion effects for face and non-face stimuli when trained with an equal number of images per category (e.g., when trained to identify the same number of human faces and birds of the same species). That is, the models showed an expertise inversion effect, not a face specific inversion effect. This contradicts the bulk of current empirical evidence showing that humans exhibit a greater inversion effect for faces compared to other categories even when they are expert at the other category. To reconcile these findings with the modelling work, Yovel et al. (2022) argue that bird watchers are more expert at human faces compared to birds, and this is why they show larger face inversion effects. Future work may well support this hypothesis, and if so, it would provide a good example of DNNs explaining important psychological data. However, as it stands, the DNN results are inconsistent with most psychological data.

This is not to say that there are no examples of DNNs doing a good job at accounting for the results from controlled experiments. For instance, **Anderson et al**. describe the results of Storrs et al. (2021) who identified conditions in which DNNs do and do not replicate illusions of gloss in humans. They found that unsupervised but not supervised learning produced human-like results and suggest unsupervised learning may play a similar role in humans. Similarly**, Op de Beeck & Bracci** describe the controlled studies by Kubilius et al (2016) showing that DNNs trained on ImageNet are sensitive to many of the non-accidental features described by Biederman (1987), a finding we found surprising but subsequently replicated in unpublished work.

However, these successes are, in our view, the exception, not the rule. A combination of relying so heavily on uncontrolled prediction-based studies, a bias against falsification in controlled studies, and selectively characterizing results to emphasize DNN-human similarities is not the way forward to advancing our understanding of human vision.

The same issues apply when large language models are also frequently compared to human language. In the target article we gave the example of Caucheteux et al. (2022) making strong conclusions about human language despite the fact that the DNNs accounted for about approximately .004 of the BOLD variance in response to spoken sentences. Similarly, Schrimpf et al. (2021) report that transformer models predict nearly 100% of explainable variance in neural responses to written sentences and suggest that "a computationally adequate model of language processing in the brain may be closer than previously thought". However, the strong claims from the article are undermined from data reported in the appendices. From Appendix S1 one finds out that the explainable variance is between 4-10% of the overall variance in three of the four datasets they analyze, and from the Appendix section "SI-1 – Language specificity", we find out that DNNs not only predict brain activation of language areas, but also nonlanguage areas, and in some analyses, the predictions are numerically larger for non-language areas. Rather than providing evidence that these models process language like humans, the correlations may be more akin to the spurious correlation observed between mouse brain activations and cryptocurrency markets (Meijer, 2021).

Furthermore, as noted by **Houghton et al.,** when a child is learning to speak, it is unlikely that she is focusing on predicting the next word. Rather, it seems likely that she is trying to communicate thoughts and desires. That is, these models learn to produce well-formed syntactic sentences when trained on arguably the wrong objective function. Similarly, these DNNs do not appear to share human-like inductive biases in learning languages, what **Bever et al**. call a universal grammar. These innate properties of humans allow the child to learn languages with many orders of magnitude less training than DNNs (human learning must be compatible with the poverty of the stimulus constraint), and at the same time, limits the types of languages that the human language system acquires (unlike language learning in DNNs; Mitchell & Bowers, 2020). In our view, research with DNNs in the domain of language provides another example that good predictions in uncontrolled studies provide little evidence that DNNs rely on human-like representations, processes, or even objective functions.

We do agree with **Houghton et al.** that it can be useful to compare language in DNNs and humans to explore the capacities of DNNs that do not have any language-specific learning mechanism. But at present, not only do the learning objectives and learning constraints seem wildly different in the two systems, but also, the performance of fully trained models "sharply diverges" from humans in controlled experiments (Huang et al., 2023).

R7 The Brain-Score neuroconnectionists

Before concluding, we thought it would be worthwhile to focus on the commentaries by **DiCarlo et al**. (7 authors) and **Golan et al.** (15 authors). Many of these authors have been amongst the most vocal in highlighting DNN-human similarities, and in both comments, they are describing agendas for how to push the field forward.

Perhaps most surprising for us**, DiCarlo et al.** do not even attempt to address the core problem with prediction-based studies used in Brain-Score, namely, predictions of observational datasets might be mediated by confounds. Instead, they mischaracterize our views regarding benchmarks, writing:

> Bowers et al. eschew community-transparent suites of benchmarks yet they imply an alternative notion of vision model evaluation, which is somehow not a suite of benchmarks… we see no alternative to support advances in models of vision other than an open, transparent and community-driven way of model comparison.

Where DiCarlo et al. get the impression that we are opposed to "open, transparent and community-driven way of model comparison" is beyond us. Rather, we caution against prediction-based studies and endorse controlled experiments to assess models, including image computable DNNs. Indeed, we are building our own (open, transparent, and community-driven) evaluation suite, that we call *MindSet,* that will make it easy for researchers to assess image computable DNNs against key findings in psychology (Biscione et al., 2023). MindSet facilitates the testing of DNNs across a series of controlled psychological experiments, each of which tests a specific hypothesis regarding how DNNs process and represent information.

The authors also report on an upcoming update on Brain-Score, with the inclusion of a controlled study by Baker and Elder (2022). They note that some DNN vision models tested on this dataset are within the noise ceiling of human data. It will be interesting to see these results given that Baker and Elder reported that VGG19, ResNet50, CorNET, and a visual transformer all failed to capture human results, writing:

> "Our configural manipulation reveals an enormous difference in how humans and networks recognize the objects: while humans rely profoundly on configural cues, networks do not".

Regardless of how current DNNs perform on this specific dataset, we welcome the introduction of controlled studies to the Brain-Score benchmark. But if the authors of Brain-Score modify their benchmark to assess the results of controlled experiments, they will need to assess models in terms of how well they explain the impact of independent variables that test specific hypotheses rather than rank models by their overall prediction accuracy.

DiCarlo et al. also defend their claim that DNNs are the current leading models of human ventral visual processing and write: "Bowers et al. critique ANN models without offering a better alternative: they imply that better models exist or should exist, but do not elaborate on what those models are”. They set the bar quite low for “best” given that current DNNs do extremely poorly in predicting the results of experiments that manipulate independent variables and provide little insight into how humans identify the objects included in current behavioral and brain benchmark studies. But in any case, we have detailed a long list of alternative models in Section 6.1 in the Target Article in Section R3 in our response. In our view, these non-image computable models have provided more insight into human vision thus far. Still, going forward, we do think it is important to try to build image computable DNNs that do account for controlled studies, and in parallel, pursue alternative modelling approaches.

Golan et al. describe a progressive Lakatosian research programme they call “neuroconnectionism” (Doerig et al., 2022) that generates a rich variety of falsifiable hypotheses and advances through model comparison. They note that neuroconnectionism itself is best thought of as a computational language that cannot be falsified and that a failure of a specific DNN does not amount to a refutation of neural network models in general. The problem with this is that no one claims that a rejection of a specific model amounts to a falsification of DNNs in general, and no one rejects modelling as a core method for advancing science. They are mounting a defence against an imaginary critique (as do other commentators, as noted in Section R2). Our criticism with neuroconnectionism is that current claims regarding DNN-human similarity are grossly overstated because researchers rely too heavily on uncontrolled prediction-based studies and avoid severe testing of their hypotheses. When the right methods are employed – namely, controlled experiments as used in virtually all other areas of science -- models account for few empirical findings of interest to vision researchers.

Unlike **DiCarlo et al.**, **Golan et al.** do note some of the advantages of controlled experiments and briefly touch on the limitations of uncontrolled prediction-based studies, writing:

> “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… 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.”

We do not find this line of argument persuasive. Yes, controlled studies are biased in the sense that they are driven by theoretical assumptions, but the unstated (and unknown) assumptions in uncontrolled studies do not avoid biased results. For example, the image datasets used in Brain-Score (see Figure 2) are not “neutral” and different results are obtained in other datasets (**Xu & Vaziri-Pashkam**). And what does it mean to claim that observational studies with naturalistic images promise to evaluate many models, and at the same time, note that this approach risks inconclusive results? Indeed, predictions made from naturalistic images taken from observational studies are, by their very nature, ambiguous as there are many potential confounds that can lead models to make predictions on the basis of short-cuts and confounds (Dujmović et al., 2023; Geirhos et al., 2020).

Furthermore, what does it mean to design tests that fall in-between observational and controlled studies? An experiment either does or does not manipulate independent variables designed to test hypotheses and rule out confounds. If the point is that it is important to work with image datasets that vary in their degree of complexity and naturalism, it remains the case that controlled experiments need to be run on all types of stimuli. Indeed, **Golan et al**. cite the discovery of texture bias and adversarial susceptibility as two examples of shortcomings of DNNs that have led to improvements. Putting aside the fact that current DNNs show almost none of the features of human shape processing and there are still no solutions to adversarial images, these limitations were both identified using controlled experiments that rely on complex but unnatural stimuli. **Golan et al.** do not identify any insights that have derived from uncontrolled studies.

Golan et al. also caricature psychology, writing: “Traditional psychological experiments are designed to test verbally defined theories”. In fact, controlled experiments have been used to assess computational models in psychology long before the invention of AlexNet (e.g., Grossberg, 1967; Hummel & Biederman, 1992; Medin & Schaffer, 1978; Ratcliff & McKoon, 2008; Rescorla & Wagner, 1972; Shepard, 1987). This general lack of regard for formal models and results in psychology (not to mention the lack of regard for verbal theories) is impeding progress in characterizing DNN-human similarities and building better models of vision and the brain more generally. Indeed, this common and unwarranted attitude towards psychology partly motivated us to write the target article in the first place.

Golan et al. also defend the claim that DNNs are the “best models” of human vision, writing:

> 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.

Here perhaps we should take the advice of **Tarr** and appreciate this is more marketing than a scientific statement.

R8. Conclusions

Human vision is both weird and wonderful. Over a century of research has identified a wide variety of properties of the human visual system that are both surprising and fundamental. To give only the most cursory of overviews, the following findings should play a central role in theory and model building. The input to our visual system is degraded due to a large blind spot and an inverted retina with light having to pass through multiple layers of retinal neurons, axons and blood vessels before reaching the photoreceptors. Nevertheless, we are unaware of the degraded signals due to a process of actively filling in missing signals in early visual cortex (e.g., Grossberg, 2003; Ramachandran, & Gregory, 1991). We have fovea that support high-acuity colour vision for only about 2 degrees of visual angle (about the size of a thumbnail at arm’s length). Nevertheless, we have the subjective sense of a rich visual experience across a much wider visual field because we move our eyes approximately 3 times per second (Rayner, 1978), with the encoding of visual inputs suppressed during each saccade (Matin, 1974), and the visual system somehow integrating inputs across fixations (Irwin, 1991). At the same time, we can identify multiple objects in scenes following a single fixation (Biederman, 1972), with object identification taking approximately 150 ms (Thorpe et al., 1996) - too quick to rely on recurrence. We are also blind to major changes in a scene as revealed by change blindness (Simons, & Levin, 1997) and have a visual short-term memory of approximately four items (Cowan, 2001). Our visual system organizes image contours by various Gestalt rules to separate figure from ground (Wagemans et al., 2012) and organize contours to build representations of object parts (Biederman, 1987). Objects are encoded in terms of their surfaces, parts, and relations between parts to build 3D representations relying on monocular and binocular inputs (Biederman, 1987; Marr, 1982; Nakayama, & Shimojo, 1992). Colour, form, and motion processing are factorized to the extent that it is possible to be cortically colour blind (Cavanagh et al., 1998), or suffer motion blindness where objects disappear during motion but are visible and recognizable while static (Zeki, 1991), or show severe impairments with object identification while maintaining the ability to reach and manipulate objects (Goodale, & Milner, 1992). Participants can even classify objects while denying seeing them (**Koculak & Wierzchon**). Our visual system manifests a wide range of visual, size, and shape constancies to estimate the distal properties of the world independent of the lighting and object pose, and we suffer from size, colour and motion illusions that reflect the very mechanisms that serve the building of these distal representations from the proximal image projected onto our retinas. These representations of distal stimuli in the world support a range of visual tasks, including object classification, navigation, grasping, and visual reasoning. All this is done with spiking networks composed of neurons with a vast range of morphologies that vary in ways relevant to their function, with architectures constrained by evolution and biophysics.

All of this and much more needs to be explained, and various modelling approaches are warranted. We agree with the commentators that one valuable approach is to keep working with current image computable DNNs while altering the tasks they solve, the data they are fed, their objective functions, learning rules, and architectures. Perhaps DNNs will converge with the biological solutions in some important respects. Whether DNNs will "automagically" (**Xu & Vaziri-Pashkam**) converge on many of these solutions when trained on the right tasks and data, however, is far from certain, and in our view, it is a mistake to put all our eggs in this one basket. Whatever approach one adopts, the current methods of evaluating DNN-human correspondences needs to change. This focus of prediction-based studies and a focus on looking for similarities and downplaying discrepancies is not going to much advance our understanding of vision and the brain more generally.

References

Anciukevicius, T., Fox-Roberts, P., Rosten, E., & Henderson, P. (2022). Unsupervised Causal Generative Understanding of Images. *Advances in Neural Information Processing Systems*, *35*, 37037-37054.

Baker, N., & Elder, J. H. (2022). Deep learning models fail to capture the configural nature of human shape perception. *Iscience*, *25*(9), 104913.

Baker, N., Garrigan, P., & Kellman, P. J. (2021). Constant curvature segments as building blocks of 2D shape representation. *Journal of Experimental Psychology: General*, *150*(8), 1556-1580.

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.

Biederman I. (1972) Perceiving real-world scenes. *Science, 177*, 77–80.

Biederman, I. (1987). Recognition-by-Components: a theory of human image understanding. *Psychological Review, 94,* 115-147.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, 103, 597-600.

Biscione, V., & Bowers, J. S. (2021). Convolutional neural networks are not invariant to translation, but they can learn to be. *The Journal of Machine Learning Research*, *22*, 10407-10434.

Biscione, V., & Bowers, J. S. (2022). Learning online visual invariances for novel objects via supervised and self-supervised training. *Neural Networks, 150*, 222-236.

Bowers, J.S. (2022). Researchers comparing DNNs to brains need to adopt standard Methods of Science. Invited workshop talk at *Neural Information Processing Systems*, New Orleans.

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, *37*, 54-115.

Caucheteux, C., Gramfort, A., & King, J. R. (2022). Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports, 12,* 1-10.

Cavanagh, P., Hénaff, M. A., Michel, F., Landis, T., Troscianko, T., & Intriligator, J. (1998). Complete sparing of high-contrast color input to motion perception in cortical color blindness. *Nature neuroscience, 1*, 242-247.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-114.

Culp, L., Sabour, S., & Hinton, G. E. (2022). Testing GLOM's ability to infer wholes from ambiguous parts. *arXiv preprint arXiv:2211.16564*.

Da Silva, L. E. B., Elnabarawy, I., and Wunsch, D. C. II. (2019). A survey of adaptive resonance theory neural network models for engineering applications. *Neural Networks, 120*, 167-203

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.

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*.

Doumas, L. A., Puebla, G., Martin, A. E., & Hummel, J. E. (2022). A theory of relation learning and cross-domain generalization. *Psychological Review*,129, 999-1041.

de Vries, J. P., Akbarinia, A., Flachot, A., & Gegenfurtner, K. R. (2022). Emergent color categorization in a neural network trained for object recognition. *ELife, 11*. https://doi.org/10.7554/eLife.76472

Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C. P., Lerchner, A., & Higgins, I. (2019). Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv*:1905.12614.

Dujmović, M., Bowers, J. S., Adolfi, F., & Malhotra, G. (2022). Some pitfalls of measuring representational similarity using Representational Similarity Analysis. *bioRxiv*, 2022-04.

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., ... & Houlsby, N. (2023). Scaling vision transformers to 22 billion parameters. *arXiv preprint* arXiv:2302.05442.

Evans, B. D., Malhotra, G., & Bowers, J. S. (2022). Biological convolutions improve DNN robustness to noise and generalisation. *Neural Networks*, *148*, 96-110.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193-202.

Francis, G., Manassi, M., & Herzog, M. H. (2017). Neural dynamics of grouping and segmentation explain properties of visual crowding. *Psychological Review, 124*, 483-504.

Garg, A. K., Li, P., Rashid, M. S., & Callaway, E. M. (2019). Color and orientation are jointly coded and spatially organized in primate primary visual cortex. *Science, 364*(6447), 1275–1279. https://doi.org/10.1126/science.aaw5868

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*, 665-673.

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.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

George, D., Lehrach, W., Kansky, K., Lázaro-Gredilla, M., Laan, C., Marthi, B., ... & Phoenix, D. S. (2017). A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science, 358*(6368).

George, D., Lazaro-Gredilla, M., Lehrach, W., Dedieu, A., & Zhou, G. (2020). A detailed mathematical theory of thalamic and cortical microcircuits based on inference in a generative vision model. *Biorxiv*, 2020-09.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*, 20-25.

Goodale, M. A., & Milner, A. D. (2023). Shape perception does not require dorsal stream processing. *Trends in Cognitive Sciences, 27, 333-334. DOI:https://doi.org/10.1016/j.tics.2022.12.007*

Grossberg, S. (1967). Nonlinear difference-differential equations in prediction and learning theory. *Proceedings of the National Academy of Sciences*, *58*, 1329-1334.

Grossberg S (2003) Filling-In the Forms: Surface and Boundary Interactions in Visual Cortex. In: *Filling-In*, pp 13–37. New York, NY: Oxford University Press

Grossberg, S. (2014). How visual illusions illuminate complementary brain processes: illusory depth from brightness and apparent motion of illusory contours. *Frontiers in Human Neuroscience*, *8*, 854.

Grossberg, S. (2021). *Conscious mind, resonant brain: how each brain makes a mind*. Oxford University Press.

Guest, O., & Martin, A. E. (2023). On Logical Inference over Brains, Behaviour, and Artificial Neural Networks. *Computational Brain & Behavior*. https://doi.org/10.1007/s42113-022-00166-x

Hermann, K. L., 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.

Hinton, G. (1979). Some demonstrations of the effects of struc- tural descriptions in mental imagery. *Cognitive Science, 3*, 231–250.

Hinton, G. (2022). How to represent part-whole hierarchies in a neural network. *Neural Computation*, 35, 413–452..

Huang, K., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2023, April 21). Surprisal does not explain syntactic disambiguation difficulty: evidence from a large-scale benchmark. https://doi.org/10.31234/osf.io/z38u6

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*, 106-152.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual cognition*, *8*, 489-517.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*, 480-517.

Hummel, J. E., & Stankiewicz, B. J. (1996). An architecture for rapid, hierarchical structural description. *Attention and performance XVI: Information integration in perception and communication*, 93-121.

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology, 23*, 420-456.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi. lingbuzz/007190

Kim, B., Reif, E., Wattenberg, M., Bengio, S., & Mozer, M. C. (2021). Neural networks trained on natural scenes exhibit gestalt closure. Computational Brain & Behavior, 4,

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*, 251-26.e1003915.

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, *12*, e1004896.

Izhikevich E. M. (2004) Which model to use for cortical spiking neurons? IEEE Transactions on Neural Networks Volume 15 Issue 5 September 2004 pp 1063–1070 https://doi.org/10.1109/TNN.2004.832719

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, *240*, 740-749.

Malhotra, G., Dujmović, M., Hummel, J., & Bowers, J. S. (in press). Human shape representations are not an emergent property of learning to classify objects. *Journal of Experimental Psychology: General*.

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin, 81*, 899–917

Marr, D. (1982) *Vision: A computational investigation into the human representation and processing of visual information*. MIT Pres

Mayo, D. G. (2018*). Statistical inference as severe testing.* Cambridge: Cambridge University Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207.

Meijer, G. (2021). Neurons in the mouse brain correlate with cryptocurrency price: a cautionary tale. *Peer Community Journal*, *1, e29*.

Messina, N., Amato, G., Carrara, F., Gennaro, C., & Falchi, F. (2021). Solving the same-different task with convolutional neural networks. *Pattern Recognition Letters, 143,* 75-80.

Mitchell, J., & Bowers, J. (2020, December). Priorless recurrent networks learn curiously. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5147-5158).

Mocz, V., Jeong, S., Chun, M., & Xu, Y. (2023). Representing Multiple Visual Objects in the Human Brain and Convolutional Neural Networks. *bioRxiv*, 2023-02.

Mocz, V., Vaziri-Pashkam, M., Chun, M., & Xu, Y. (2022). Predicting Identity-Preserving Object Transformations in Human Posterior Parietal Cortex and Convolutional Neural Networks. *Journal of Cognitive Neuroscience*, *34*, 2406-2435.

Montero, M. L., Ludwig, C. J., Costa, R. P., Malhotra, G., & Bowers, J. (2021, May). The role of disentanglement in generalisation. In *International Conference on Learning Representations*.

Montero, M., Bowers, J., Ponte Costa, R., Ludwig, C., & Malhotra, G. (2022). Lost in Latent Space: Examining failures of disentangled models at combinatorial generalisation. *Advances in Neural Information Processing Systems*, *35*, 10136-10149.

Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science, 257*, 1357-1363.

Newell, A. (1973). You can’t play 20 questions with nature and win: Projective comments on the papers of this symposium. In Chase, W. G. (Ed.). 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.

Puebla, G., & Bowers, J. S. (2022). Can deep convolutional neural networks support relational reasoning in the same-different task?. *Journal of Vision, 22*, 11. doi:https://doi.org/10.1167/jov.22.10.11

Puebla, G., & Bowers, J. S. (2023). The role of object-centric representations, guided attention, and external memory on generalizing visual relations. *arXiv preprint arXiv:2304.07091*.

Qin, Y., Frosst, N., Sabour, S., Raffel, C., Cottrell, G., & Hinton, G. (2019). Detecting and diagnosing adversarial images with class-conditional capsule reconstructions. *arXiv preprint arXiv:1907.02957*.

Ramachandran, V. S., & Gregory, R. L. (1991). Perceptual filling in of artificially induced scotomas in human vision. *Nature*, *350*, 699-702.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, *20*, 873-922.

Rawski, J. & Baumont, J. (2023). Modern Language Models Refute Nothing. Lingbuzz Preprint.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*, 618-660.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory, 2*, 64-69.

Rich, P., de Haan, R., Wareham, T., & van Rooij, I. (2021). How hard is cognitive science?. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43, No. 43).

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, *30*.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*, e2105646118.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances, 8*, eabm2219.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences, 1*, 261-267.

Simmons J. P., Nelson L. D., Simonsohn U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience, 33*, 2044-2064

Tang, K., Chin, M., Chun, M., & Xu, Y. (2022). The contribution of object identity and configuration to scene representation in convolutional neural networks. *Plos one*, *17*(6), e0270667.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520-522.

Tsvetkov, C., Malhotra, G., Evans, B. D., & Bowers, J. S. (2023). The role of capacity constraints in Convolutional Neural Networks for learning random versus natural data. *Neural Networks*, *161*, 515-524.

Vannuscorps, G., Galaburda, A., & Caramazza, A. (2021). The form of reference frames in vision: The case of intermediate shape-centered representations. *Neuropsychologia*, *162*, 108053.

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, 1–2. https://doi.org/10.1038/s44159-022-00031-5

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological Bulletin, 138*, 1172.

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

Xu, Y., & Vaziri-Pashkam, M. (2022). Understanding transformation tolerant visual object representations in the human brain and convolutional neural networks. *Neuroimage*, *263*, 119635.

Yamins, D. L., 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, 111*, 8619-8624.

Yovel, G., Grosbard, I., & Abudarham, N. (2022). Computational models of perceptual expertise reveal a domain-specific inversion effect for objects of expertise. PsyXiv.

Zeki S. (1991)/ Cerebral akinetopsia (visual motion blindness). A Review. *Brain*, 114, 811–824.

Zhang, H., Zhang, Y. F., Liu, W., Weller, A., Schölkopf, B., & Xing, E. P. (2022). Towards principled disentanglement for domain generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8024-8034).

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications, 10*, 1-9.