A comparative biology approach to DNN modeling of vision: A focus on differences, not similarities

Ben Lonnqvist Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Alban Bornet Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Adrien Doerig Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands

Michael H. Herzog Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Deep neural networks (DNNs) have revolutionized computer science and are now widely used for neuroscientific research. A hot debate has ensued about the usefulness of DNNs as neuroscientific models of the human visual system; the debate centers on to what extent certain shortcomings of DNNs are real failures and to what extent they are redeemable. Here, we argue that the main problem is that we often do not understand which human functions need to be modeled and, thus, what counts as a falsification. Hence, not only is there a problem on the DNN side, but there is also one on the brain side (i.e., with the explanandum—the thing to be explained). For example, should DNNs reproduce illusions? We posit that we can make better use of DNNs by adopting an approach of comparative biology by focusing on the differences, rather than the similarities, between DNNs and humans to improve our understanding of visual information processing in general.

The explanans

Deep neural networks (DNNs) have revolutionized the field of computer vision, reaching or exceeding human performance in object recognition tasks (LeCun, Bottou, Bengio, & Haffner, 1998; Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015; He, Zhang, Ren, & Sun, 2015). This excellent performance and the analogy between DNNs and the primate visual cortex have caused a fierce discussion about the use of DNNs as neuroscientific models of the brain (e.g., Rosenholtz, 2017; VanRullen, 2017; Majaj & Pelli, 2018; Kubilius, 2018; Cichy & Kaiser, 2019; Kietzmann, McClure, & Kriegeskorte, 2019a; Richards et al., 2019; Firestone, 2020; Griffiths, 2020; Lindsay, 2020; Saxe, Nelli, & Summerfield, 2020; see also DiCarlo, Zoccolan, & Rust, 2012). This debate is about the explanans—that is, how good DNNs are as *models* for human brain processing and behavior.

It is first important to make a distinction among the different types of models (Cichy & Kaiser, 2019). We distinguish three general groups: functional models, which attempt to capture the most important behavioral characteristics of a system; mechanistic models, which attempt to recreate the underlying implementation of the system itself; and replica models, which attempt to do both as accurately as possible. For example, a replica model of the human brain might be an in silico model, where all details, neural spikes, and mRNA coding, for example, are captured. The Human Brain Project (Markram et al., 2011; Amunts, Ebell, Muller, Telefont, Knoll, & Lippert, 2016) is an example that aims to create such a replica. With such a model, neural processes could be studied as if they were factually human brain processes. For example, one may knock out the dopaminergic system and study what neural functions it contributes to. Whether such models are practically possible is an open question. When we use

Citation: Lonnqvist, B., Bornet, A., Doerig, A., & Herzog, M. H. (2021). A comparative biology approach to DNN modeling of vision: A focus on differences, not similarities. *Journal of Vision*, *21*(10):17, 1–10, https://doi.org/10.1167/jov.21.10.17.

https://doi.org/10.1167/jov.21.10.17 Received February 26, 2021; published September 22, 2021 ISSN 1534-7362

the term "model," we do not consider replica models unless mentioned explicitly.

Typically, models attempt to describe only some aspects of a system. These models abstract away certain aspects either by condensing complex processes into simpler components or by ignoring them entirely. A model of a neuron may ignore the complex molecular machinery in the nucleus of the neuron without sacrificing the predictive power of certain aspects. Such a model is functional, as it does not explain (nor does it attempt to) the mechanisms by which the neural responses are generated.

Mechanistic models, as opposed to functional models, attempt to describe the mechanisms by which certain behaviors of the system arise. A mechanistic model of a neuron, unlike a functional one, would not ignore the complex molecular biology underlying the function of the neuron; instead, it would attempt to explain how those processes contribute to the behavior of the neuron but not necessarily the entire system.

The debate about the use of DNNs as models of human vision centers on several arguments, of which the following three are of most importance.

First and foremost, DNNs can predict neural activation in primate visual cortex better than other preceding models (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins, Hong, Cadieu, Solomon, Seibert, & DiCarlo, 2014; Kubilius, Schrimpf, Nayebi, Bear, Yamins, & DiCarlo, 2018; but see Eberhardt, Cader, & Serre, 2016; Xu & Vaziri-Pashkam, 2021). This fact is strong evidence that DNNs share something crucial in common with human visual processing that other traditional models do not (whether it be general similarities of architecture, optimization, or something else).

Second, DNNs have shown great promise for modeling human psychophysical tasks, such as image recognition (e.g., Geirhos, Rubisch, Michaelis, Bethge, Wichmann, & Brendel, 2018; Su, Vargas, & Kouichi, 2019; Geirhos, Meding, & Wichmann, 2020; Geirhos, Narayanappa, Mitzkus, Thieringer, Bethge, Wichmann, & Brendel, 2021) or crowding, a breakdown of object recognition in the presence of surrounding objects (Volokitin, Roig, & Poggio, 2017; Doerig, Bornet, Choung, & Herzog, 2020a; Lonnqvist, Clarke, & Chakravarthi, 2020). However, even though DNNs show close to human-like object recognition performance, their processing can be highly different than that of humans. For example, ImageNet-trained DNNs prefer textural information rather than the shape-based information that humans prioritize (Geirhos et al., 2018). The trial-by-trial performance of DNNs in perceptual tasks is also consistently different than that of humans (Geirhos et al., 2020; Geirhos et al., 2021). Likewise, although on a category-to-category basis the response patterns of DNNs may appear similar to those of humans, the specific images on which DNNs make misclassifications are often different from the images on which humans make misclassifications (Geirhos et al., 2020). This suggests systematic differences in categorization ability. Even the specifically brain-inspired recurrent CORnet-S shows response patterns that are similar to those of other DNNs and dissimilar to human response patterns (Rajalingham, Issa, Bashivan, Kar, Schmidt, & DiCarlo, 2018; Geirhos et al., 2020). This indicates that the function they compute to solve a task, regardless of architectural specifics, remains largely different from that of humans. Hence, even though performance of DNNs and humans may be similar, the computation underlying the performance may be very different.

Finally, DNNs are prone to overfitting and adversarial attacks (e.g., Goodfellow, Shlens, & Szegedy, 2014; Su et al., 2019; Dujmović, Malhotra, & Bowers, 2020). An interesting example is the one-pixel attack (Su et al., 2019), whereby changing a single pixel in an image can cause a DNN to misclassify the image. These attacks may depend on the dataset, however, rather than on the model architecture (Ilyas, Santurkar, Tsipras, Engstrom, Tran, & Madry, 2019; Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021). This implies that DNNs trained with vulnerable datasets, such as ImageNet and CIFAR, are not good representatives of the potential of DNNs to serve as models of human vision, as humans are not vulnerable to such adversarial attacks. There is ongoing research as to what defenses can be employed to make DNNs robust to adversarial attacks (for reviews, see Xu, Ma, Liu, Deb, Liu, Tang, & Jain, 2019; Machiraju, Choung, Frossard, & Herzog, 2021), and recent research has demonstrated large improvements in robustness (e.g., Dapello et al., 2020; Radford et al., 2021).

Taken together, there are clear shortcomings of DNNs as models of the human system. The question is how serious these shortcomings are. Firestone (2020) pointed out that human–DNN comparisons may often not be "fair." Indeed, human–DNN comparisons have little meaning if the methods by which humans and DNNs are tested are dissimilar. Firestone argues that the process for making these comparisons fair is threeway: One must limit DNNs like humans, limit humans like DNNs, and align tasks to consider the species performing them. It remains unclear whether DNNs generally fail to exhibit human-like effects in visual phenomena or whether this dissimilarity is caused by unfair comparisons. As Firestone (2020) argued, suppose a DNN is optimized to exploit image-level textural information due to a bias in a dataset. It would then not be fair to say that, because this DNN places more emphasis on textural information (as opposed to, for example, shape information), DNNs generally process information differently than humans do (Geirhos et al., 2018). Given these conflicting arguments and studies, the status of DNNs as models

of information processing in the human visual system remains unclear.

The explanandum

In the last section, we reviewed the discussion of whether or not DNNs describe human brain processing and behavior well; that is, the focus was on the explanans. In this section, we argue that there is an issue with the explanandum, as well; we do not understand brain processing well enough to determine what is crucial and what is not. Hence, we cannot know what can or should be abstracted away by a model and what cannot or should not be. In other words, we often do not know when a model is falsified and when it is not. If we do not know what phenomenon we want to explain, the question of whether or not DNNs are good at explaining the phenomenon is irrelevant.

The neural explanandum

Here, we describe cases where we know little about neural processing. First, it is a great success that DNNs can predict neural responses well in terms of correlations of DNN and primate neural spike rates. However, it is still a mystery what the neural code of the brain is, and perhaps it is possible that specific spike rates are less crucial than thought; thus, these correlations may only pick up some epiphenomenal aspects of brain processing. Second, we argue that metrics such as the Brain-Score (Schrimpf, 2018) may not provide sufficient constraints for meaningful model selection on a neural level due to the low resolution of the data on which they are computed.

When modeling the human visual system and judging success rates based on a metric, we rank different explanans (explanations) of the underlying explanandum (in this case, the human visual system) according to this metric. The metric (e.g., Brain-Score; Schrimpf, 2018) serves as a ranking of the explanans for that explanandum. In other words, we are explaining performance on a metric using our models and hoping that improvements in explaining the metric generalize to improvements in explaining the underlying system (the brain).

One problem with this approach is that current metrics do not allow model selection. For example, the Brain-Score benchmark (Schrimpf et al., 2018; Schrimpf, Kubilius, Lee, Ratan Murty, Ajemian, & DiCarlo, 2020) combines a number of neural predictivity scores, such as correlations between neural activity in DNN layers and the visual cortex (V1, V2, V4) and inferior temporal (IT) cortex, as well as correlations between behavioral measures of DNNs and humans. Interestingly, many of the top-ranking models on the Brain-Score are architecturally substantially different, yet show similar performance in terms of Brain-Score. For example, simulating the human primary visual cortex at the front of a convolutional network (Dapello et al., 2020) barely improves Brain-Score compared with other models, such as the early brain-inspired VGG-19 (Simonyan & Zisserman, 2015); a recurrent brain-inspired model, CORnet-S (Kubilius et al., 2018), or other convolutional networks (He et al., 2015; Huang, Liu, Van Der Maaten, & Weinberger, 2017).

This is problematic because these metrics fail to adequately distinguish among substantially different model architectures. Although there are properties of DNNs that generalize to the human visual system according to the Brain-Score metric (such as convolutions), direct modeling efforts have not been able to make substantial progress in explaining the human visual system beyond this. If a metric fails to discriminate among models that are recurrent (Kubilius et al., 2019), fully feedforward (Simonyan & Zisserman, 2015), residual (He et al., 2015; Huang et al., 2017), or other (e.g., Kolesnikov, Zhai, & Beyer, 2019; Dapello et al., 2020), then perhaps the metric is insufficient. The main problem is that model failure is not clearly defined in this case; either the Brain-Score metric or the methodology with which a model is evaluated on it fails to distinguish among what we would think of as fundamentally different types of model architectures.

One possible cause of the problem is the low resolution of neural imaging data. In Brain-Score, for example, correlation in V4 and IT is computed on the activity of 88 V4 neurons and two datasets of 168 IT neurons and on five 96-electrode Utah arrays in IT, respectively (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013; Majaj, Hong, Solomon, & DiCarlo, 2015; Schrimpf et al., 2018; Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019). In a recent article, Xu and Vaziri-Pashkam (2021) demonstrated (in a functional magnetic resonance imaging study) that higher neural resolution data can allow for substantially better model discrimination. This suggests that more work is needed on collecting higher resolution datasets to allow better model selection based on neural metrics.

Similar results are found in human studies. Predicting the human brain activity of one human with the neural activity of another in different modalities, such as vision (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004) and social interaction (Dumas, Nadel, Soussignan, Martinerie, & Garnero, 2010), is not much more precise than predicting human neural data with DNNs. This has been formalized as a "noise ceiling," an upper bound on the linear predictability of the human brain derived from predictability of the human brain from other primates, and recent evidence suggests this ceiling may

have been reached (Khaligh-Razavi & Kriegeskorte, 2014; Storrs, Khaligh-Razavi, & Kriegeskorte, 2020a; Storrs, Kietzmann, Walther, Mehrer, & Kriegeskorte, 2020b; but see Xu & Vaziri-Pashkam, 2021).

Current metrics such as the Brain-Score are not sufficient to understand what is relevant for the human visual system, as they mask important aspects about how a task can be performed in different ways. In this case, several architectures can perform similarly along the Brain-Score metric, and hence it is unclear what architectural parts of DNNs are crucial in this respect. For this reason, it remains unclear which neural code is used by the visual cortex.

The psychophysical explanandum

Just as it is difficult to find which neural activities to model, deciding which behavioral characteristics to model is not straightforward, either. Here, we discuss the importance of being explicit about the phenomena we wish to model and about which results would falsify a model.

Scientists may have reasons to think some phenomena of human vision are idiosyncratic, but others are ubiquitous and crucial for understanding visual processing. Studying seemingly idiosyncratic phenomena using traditional models often requires careful thought about how to phrase the phenomenon in a way that allows for the model to clearly display these effects. In contrast, DNNs allow us to study either type of phenomenon with relative ease compared with traditional types of models. This is because model formulation itself may involve as little as selecting an existing DNN architecture, and a dataset to train it with. Whether or not the underlying phenomenon being modeled is ubiquitous is not important from the point of view of the model. DNNs allow us to conveniently place the explanans before the explanandum because they are not generally explicitly hypothesis driven (barring superficial architectural similarities). In other words, justifying why a phenomenon (such as a visual illusion) should be modeled under the DNN framework is easily ignored.

It is therefore doubly important that we carefully justify the phenomena we model using DNNs to avoid finding patterns where there are none. In other words, let us not "blanket model" every phenomenon simply because we can.

Here, we show that it is not always clear which phenomena should be modeled and which should not. We illustrate this with two phenomena: crowding and illusions. We argue that, in the case of visual crowding, we know what we want to model and how we can validate and falsify models, but that the same cannot be said for visual illusions.

Example 1: Crowding

Crowding is the ubiquitous breakdown of object recognition in the presence of nearby flankers (Bouma, 1970; Pelli, Palomares, & Majaj, 2004; Levi, 2008; Sayim, Westheimer, & Herzog, 2008; Manassi & Whitney, 2018). In simple stimulus configurations, crowding is easily described by Bouma's law (Bouma, 1970; Pelli & Tillman, 2008): Only flankers within a certain window around the target deteriorate performance. The window size is often estimated to be half of the eccentricity of the target location; however, recent work has shown that more complex computational processes underlie crowding. Importantly, when adding flankers, object recognition improves under certain conditions (Manassi, Sayim, & Herzog, 2013; Manassi, Hermens, Francis, & Herzog, 2015; Herzog & Manassi, 2015). This phenomenon, called uncrowding, can occur even when adding flankers outside of Bouma's window. Uncrowding challenges most models of vision because more flankers can only deteriorate performance in these models. For example, a crucial operation in the early layers of DNNs is pooling information across neighboring spatial locations. More flankers diminish target signals and hence psychophysical performance.

Crowding and, relatedly, uncrowding are ubiquitous phenomena in vision, as stimuli are rarely encountered in isolation. In this respect, we argue that any successful object recognition system must cope with crowding because of its ubiquity; if a model does not produce crowding and uncrowding, it should be rejected. Study of such models could provide insight into the purpose and consequences these phenomena may carry for visual processing systems (e.g., see Doerig, Schmittwilken, Sayim, Manassi, & Herzog, 2020b, who show evidence for the importance of recurrent segmentation in visual processing).

On the other hand, crowding shows interesting phenomena such as anisotropies (Toet & Levi, 1992), whereby flankers away from fixation crowd more strongly than flankers closer to fixation. We would not require a model to explain this phenomenon, as we have in the model-building process abstracted away from heterogeneities in the photoreceptor distribution, which are believed to cause these anisotropies. Hence, crowding is a function of the visual system where we likely know what counts as a failure of a model in the context of vision and what does not. The rejection criteria for models are clear, because we likely know what we *should* model and are explicit about what it is we are *attempting* to model.

Many DNNs have failed to reproduce uncrowding, highlighting that they process visual information very differently from humans (Doerig et al., 2020a; Lonnqvist et al., 2020). However, adding an explicit

segmentation stage can remedy the models (Doerig et al., 2020b). Thus, DNNs are not rejected as models of human vision in general, but only certain types of DNNs are. In this case, we have successfully done model selection.

Example 2: Illusions

Here, we argue that visual illusions are a case where the explanandum is not clear, and because of that modeling visual illusions may be premature. We do not currently understand in most cases why illusions appear in the human visual system; for example, it is not known whether they are "bugs" of the visual system or whether they are a feature (for a discussion about veridicality of illusions, see, e.g., Braddick, 1972; Braddick, 2018; Todorović, 2018; Todorović, 2020). One may argue that DNNs need to capture visual illusions simply because they are part of human vision, but this can be said about all visual function and we return to the replica model case. Alternatively, one may argue that illusions are just idiosyncratic failures of the visual system that we should abstract away as in the case of the anisotropies of crowding.

Many studies have investigated whether and to what extent DNNs are susceptible to illusions, and they have found an array of conflicting results. For example, training DNNs on complex tasks made them susceptible to visual illusions (Mély, Linsley, & Serre, 2018; Watanabe, Kitaoka, Sakamoto, Yasugi, & Tanaka, 2018; Benjamin, Qiu, Zhang, Kording, & Stocker, 2019; Gomez-Villa, Martín, Vazquez-Corral, & Bertalmío, 2019), but not consistently so (Gomez-Villa, Martín, Vazquez-Corral, Bertalmío, & Malo, 2020). Gomez-Villa et al. (2020) showed that small DNNs do not exhibit the gradient illusion like humans but exhibit many classic illusions (such as the dungeon illusion and White's illusion). In contrast, state-of-the-art DNNs (Zhang, Zuo, Chen, Meng, & Zhang, 2017; Tao, Gao, Shen, Wang, & Jia, 2018) exhibit human-like effects in the gradient illusion but weak to no effect on others. In general, it appears difficult to find a coherent way of relating these conflicting results to human illusory processing (e.g., Baker, Erlikhman, Kellman, & Lu, 2018; Sun & Dekel, 2019; Ward, 2019). Thus, should we dismiss all models because none of them reproduces all human illusions, or are some illusions more important than others, allowing us to do model selection accordingly? We simply do not know. The problem is with the explanandum, not with the explanans.

In fact, the situation is even more complex. The lack of coherence is not limited to DNNs but also extends to human studies. Humans exhibit large individual variation in illusion strength, and performance in most illusion tasks poorly predicts performance in other illusions, even when the illusions are qualitatively similar (Grzeczkowski, Clarke, Francis, Mast, & Herzog, 2017; Cretenoud, Karimpur, Grzeczkowski, Francis, Hamburger, & Herzog, 2019; Cretenoud, Grzeczkowski, Bertamini, & Herzog, 2020). Importantly, some humans are not deceived by certain illusions at all. Thus, what should we model? The problem is not the heterogeneity of visual illusions per se (this issue can be escaped by modeling specific illusions on a case-by-case basis) but rather the fact that, because of the large degree of heterogeneity in visual illusions even within subjects, it is not yet possible to determine which illusions are crucial for understanding vision. This raises the question of which illusions need to be modeled in a successful model. There is no good answer that we know of.

The fact that there is little coherence within DNN studies and, similarly, little coherence in human studies should not be taken as evidence of the similarity of DNNs to humans in illusory perception. The real problem relates to the human brain and about how it is impossible to know which illusions (if any) are important to human vision. In summary, until we know which illusions reflect crucial and ubiquitous aspects of human visual processing (as we argued that we do in the case of visual crowding), attempting to model them in a vacuum does us little good. We are unable to place our findings into context and will find difficulty in discriminating between a successful and an unsuccessful model. It may be that the situation is similar for other visual functions, so that it is not clear what needs to be modeled and what aspects should be abstracted away.

When the explanandum becomes the explanans: An approach of comparative biology

In this section, we argue that we can learn a great deal from DNNs when we consider them as independent visual species—in other words, by considering them as visual systems in their own right rather than as models of the human visual system. The approach is akin to comparative biology research. Here, we offer some important insights we have gained from DNNs under this framework.

First, potentially the deepest insight for the vision community may have been that object recognition can occur without carefully modeling the visual system step by step along its hierarchy. Over the last half century, vision research has been split up in many subcommunities that are concerned with a variety of topics, such as shape, motion, color, and others. Even though there has been little cross-talk between these subcommunities, all of these fields unify the implicit notion that the visual system must be able to

solve all of these aspects individually for successful object recognition. DNNs have demonstrated that a system can perform well on object recognition without explicitly training on or even performing well in many of the aforementioned areas (for example, DNNs apparently solve vision without relying on shape information; Geirhos et al., 2018). This provokes the question as to what extent some of the visual processes we have studied in the last 50 years are crucial and representative of vision in general.

Second, a particular question was whether object recognition necessarily requires explicit object segmentation (Herzog, Sayim, Chicherov, & Manassi, 2015) or can occur without it (VanRullen & Thorpe, 2002). DNNs have provided a clear answer to this question in favor of the latter hypothesis (e.g., Krizhevsky et al., 2012; Simonyan & Zisserman, 2015). Here, a new question arises. Does a computational advantage exist in favor of segmentation, or is it simply a suboptimality caused by either our environment or evolution? In this respect, the failure of DNNs to exhibit human-like crowding and uncrowding is not actually a failure but rather a useful clue that may offer insights for what is crucial for which types of visual function.

Third, neuroscience often relies on a subpart coding strategy. The coding of neurons can be mapped directly onto perceptual aspects such as V1 activity to edge perception, IT activity to faces, and even mapping individual faces to neurons. Neurons of DNNs show a coding with relatively interpretable features but not ones reliably mappable to the brain (e.g., Xu & Vaziri-Pashkam, 2021). In addition, different DNNs code subparts differently (e.g., Olah, Mordvintsev, & Schubert, 2017; Geirhos et al., 2018). These aspects are not surprising, as it is known that one can use infinitely many orthonormal bases, such as Fourier or Gabor wavelets, to code for any function (i.e., representation of a stimulus). Hence, there are many ways to code subparts, and DNNs may show the extent to which details of how subparts are coded does not matter. Likewise, artificial vision can be achieved in many different ways, including a large variety of feedforward DNNs (Krizhevsky et al., 2012) or recurrent DNNs (Kietzmann, Spoerer, Sörensen, Cichy, Hauk, & Kriegeskorte, 2019b), with (e.g., Dosovitskiy et al., 2020; Radford et al., 2021) or without (e.g., Krizhevsky et al., 2012) attention and even with very different architectures, such as transformers (Dosovitskiy et al., 2020) and multilayer perceptrons (Tolstikhin et al., 2021).

Finally, DNNs often overfit and show little generalization, as revealed by, for example, adversarial examples (Goodfellow et al., 2014; Ilyas et al., 2019; Su et al., 2019; Dujmović et al., 2020; Mehrer et al., 2021). One may consider this a failure of DNNs. However, we see this fact rather as an invitation to study the question to what extent overfitting is a bug or a feature, given the large resources DNNs and human brains have (Ilyas et al., 2019). In addition, we can ask to what extent humans overfit, as well (for a study about human overfitting, see, e.g., Dubey, Agrawal, Pathak, Griffiths, & Efros, 2018). For example, human perceptual learning is very specific and can be argued to be a case of overfitting (e.g., Spang, Grimsen, Herzog, & Fahle, 2010).

Conclusions

DNN model selection is currently difficult. We have highlighted issues with regard to this on both the side of the explanandum and the side of the explanans. These issues are not necessarily ubiquitous but can potentially be overcome with further improvement of model validation methods, an increased understanding of different phenomena in the human visual system, and specificity about which phenomena are to be modeled. We argue that, as a whole, accepting DNNs as validated models of the human visual system is premature. In addition, more work is needed to understand what we want to model and what we do not; we should understand the explanandum first. As shown, successful object recognition per se is not a benchmark because it can be achieved in many ways. More fine-grained benchmarks are needed when it comes to both neural processing and psychophysics, and often it is unclear which benchmarks should be used, because we simply do not understand vision in many respects.

Until then, comparisons may be more fruitful when focusing on differences between DNNs and humans. Studying how and why it is possible to achieve a goal differently can offer insight on what is crucial for performing the task. A comparative biology approach can be used as a key step not only in understanding how DNNs function but also in understanding how visual information processing functions *in general* beyond specific species, be it humans, DNNs, or other animals such as chickens (Ciandetti & Vallortigara, 2018). We suggest that we should consider both DNNs and human vision as different subsets of visual information processing; they are different species.

This approach has told us much already; for example, we have learned that large-scale object recognition can be achieved with or without several functions or functional properties of the human visual system, such as attention, segmentation, or recurrence. Ultimately, we think that there can exist a compelling synergy between DNN modeling and neuroscience. A feedback loop of new insights gained from direct modeling (such as new hypotheses or compelling models of specific processes) can be combined with human studies and human–DNN comparative studies to

produce a rigorous body of research that facilitates an understanding of the principles underlying visual information processing in general.

Keywords: deep neural networks, modeling, comparative biology, crowding, illusions

Acknowledgments

BL and MHH were supported by a grant from the Swiss National Science Foundation (N.320030_176153; "Basics of visual processing: from elements to figures"). AD was supported by a grant from the Swiss National Science Foundation (N.191718; "Towards machines that see like us: human eye movements for robust deep recurrent neural networks"). AB was supported by the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements N.785907 (Human Brain Project SGA2) and N.945539 (Human Brain Project SGA3).

Commercial relationships: none.
Corresponding author: Ben Lonnqvist.
Email: ben.lonnqvist@epfl.ch.
Address: Laboratory of Psychophysics, Brain Mind Institute, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland.

References

Amunts, K., Ebell, C., Muller, J., Telefont, M., Knoll, A., & Lippert, T. (2016). The Human Brain Project: Creating a European research infrastructure to decode the human brain. *Neuron, 92*(3), 574–581, https://doi.org/10.1016/j.neuron.2016.10.046.

Baker, N., Erlikhman, G., Kellman, P. J., & Lu, H. (2018). Deep convolutional networks do not perceive illusory contours. *Cognitive Science*, 1310–1315.

Benjamin, A., Qiu, C., Zhang, L.-Q., Kording, K., & Stocker, A. (2019). Shared visual illusions between humans and artificial neural networks. In *2019 Conference on Cognitive Computational Neuroscience* (pp. 585–588), https://doi.org/10.32470/CCN.2019.1299-0.

Bouma, H. (1970). Interaction effects in parafoveal letter recognition. *Nature, 226*(5241), 177–178, https://doi.org/10.1038/226177a0.

Braddick, O. (1972). Review: The psychology of visual illusion and some illusions of visual psychology: The psychology of visual illusion. *Perception, 1*(2), 239–241, https://doi.org/10.1068/p010239.

Braddick, O. (2018). Illusion research: An infantile disorder? *Perception, 47*(8), 805–806, https://doi.org/10.1177/0301006618774658.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., & Solomon, E. A., …DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology, 10*(12), 1–18, https://doi.org/10.1371/journal.pcbi.1003963.

Ciandetti, C., & Vallortigara, G. (2018). Chicken – Cognition in the poultry yard. In N. Bueno-Guerra, & F. Amici (Eds.), *Field and laboratory methods in animal cognition: A comparative guide* (pp. 97–111). Cambridge, UK: Cambridge University Press.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317, https://doi.org/10.1016/j.tics.2019.01.009.

Cretenoud, A. F., Grzeczkowski, L., Bertamini, M., & Herzog, M. H. (2020). Individual differences in the Müller-Lyer and Ponzo illusions are stable across different contexts. *Journal of Vision, 20*(6):4, 1–14, https://doi.org/10.1167/jov.20.6.4.

Cretenoud, A. F., Karimpur, H., Grzeczkowski, L., Francis, G., Hamburger, K., & Herzog, M. H. (2019). Factors underlying visual illusions are illusion-specific but not feature-specific. *Journal of Vision, 19*(14):12, 1–21, https://doi.org/10.1167/19.14.12.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)* (pp. 1–15). San Diego, CA: Neural Information Processing Systems.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron, 73*(3), 415–434.

Doerig, A., Bornet, A., Choung, O. H., & Herzog, M. H. (2020a). Crowding reveals fundamental differences in local vs. Global processing in humans and machines. *Vision Research, 167*, 39–45, https://doi.org/10.1016/j.visres.2019.12.006.

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020b). Capsule networks as recurrent models of grouping and segmentation. *PLoS Computational Biology, 16*(7), e1008017, https://doi.org/10.1371/journal.pcbi.1008017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., & Unterthiner, T., …Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, arXiv:2010.11929v2 [cs.CV].

Dubey, R., Agrawal, P., Pathak, D., Griffiths, T. L., & Efros, A. A. (2018). Investigating human priors for playing video games. *arXiv*, arXiv:1802.10217v3 [cs.AI].

Dujmović, M., Malhotra, G., & Bowers, J. S. (2020). What do adversarial images tell us about human vision? *Elife, 9*, e55978.

Dumas, G., Nadel, J., Soussignan, R., Martinerie, J., & Garnero, L. (2010). Inter-brain synchronization during social interaction. *PLoS One, 5*(8), e12166.

Eberhardt, S., Cader, J., & Serre, T. (2016). How deep is the feature analysis underlying rapid visual categorization? *arXiv*, arXiv:1606.01167v1 [cs.CV].

Firestone, C. (2020). Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences, USA, 117*(43), 26562–26571.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*(7), 974–981.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of DNNs and humans by measuring error consistency. *arXiv*, arXiv:2006.16736v3 [cs.CV].

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. *arXiv*, arXiv:2106.07411v1 [cs.CV].

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained DNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv*, arXiv:1811.12231v2 [cs.CV].

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., & Bertalmío, M. (2019). Convolutional Neural Networks Can Be Deceived by Visual Illusions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12301–12309). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., & Malo, J. (2020). Color illusions also deceive DNNs for low-level vision tasks: Analysis and implications. *Vision Research, 176*, 156–174, https://doi.org/10.1016/j.visres.2020.07.010.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv*, arXiv:1412.6572v3 [stat.ML].

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences, 24*(11), 873–883, https://doi.org/10.1016/j.tics.2020.09.001.

Grzeczkowski, L., Clarke, A. M., Francis, G., Mast, F. W., & Herzog, M. H. (2017). About individual differences in vision. *Vision Research, 141*, 282–292, https://doi.org/10.1016/j.visres.2016.10.006.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*(5664), 1634–1640.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv*, arXiv:1512.03385v1 [cs.CV].

Herzog, M. H., & Manassi, M. (2015). Uncorking the bottleneck of crowding: A fresh look at object recognition. *Current Opinion in Behavioral Sciences, 1*, 86–93, https://doi.org/10.1016/j.cobeha.2014.10.006.

Herzog, M. H., Sayim, B., Chicherov, V., & Manassi, M. (2015). Crowding, grouping, and object recognition: A matter of appearance. *Journal of Vision, 15*(6):5, 1–18, https://doi.org/10.1167/15.6.5.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4700–4708). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (pp. 1–12). San Diego, CA: Neural Information Processing Systems.

Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience, 22*(6), 974–983.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*(11), e1003915.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019a). Deep neural networks in computational neuroscience. In *Oxford research encyclopedia of neuroscience*. Oxford, UK: Oxford University Press.

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019b). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences, USA, 116*(43), 21854–21863.

Kolesnikov, A., Zhai, X., & Beyer, L. (2019). Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1920–1929). Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems* (pp. 1097–1105). San Diego, CA: Neural Information Processing Systems.

Kubilius, J. (2018). Predict, then simplify. *NeuroImage, 180*, 110–111, https://doi.org/10.1016/j.neuroimage.2017.12.006.

Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N. J., & Rajalingham, R., …DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *arXiv*, arXiv:1909.06161v2 [cs.CV].

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). CORnet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, https://doi.org/10.1101/408385.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324, https://doi.org/10.1109/5.726791.

Levi, D. M. (2008). Crowding–an essential bottleneck for object recognition: A mini-review. *Vision Research, 48*(5), 635–654, https://doi.org/10.1016/j.visres.2007.12.009.

Lindsay, G. W. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience, 33*(10), 1–15, https://doi.org/10.1162/jocn_a_01544.

Lonnqvist, B., Clarke, A. D. F., & Chakravarthi, R. (2020). Crowding in humans is unlike that in convolutional neural networks. *Neural Networks, 126*, 262–274, https://doi.org/10.1016/j.neunet.2020.03.021.

Machiraju, H., Choung, O.H., Frossard, P., & Herzog, M. H. (2021). Bio-inspired robustness: A review. *arXiv*, arXiv:2103.09265v1 [cs.CV].

Majaj, N. J., Hong, H., Solomon, E. A., & DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience, 35*(39), 13402–13418.

Majaj, N. J., & Pelli, D. G. (2018). Deep learning—Using machine learning to study biological vision. *Journal of Vision, 18*(13):2, 1–13, https://doi.org/10.1167/18.13.2.

Manassi, M., Hermens, F., Francis, G., & Herzog, M. H. (2015). Release of crowding by pattern completion. *Journal of Vision, 15*(8):16, 1–15, https://doi.org/10.1167/15.8.16.

Manassi, M., Sayim, B., & Herzog, M. H. (2013). When crowding of crowding leads to uncrowding. *Journal of Vision, 13*(13):10, 1–10, https://doi.org/10.1167/13.13.10.

Manassi, M., & Whitney, D. (2018). Multi-level crowding and the paradox of object recognition in clutter. *Current Biology, 28*(3), R127–R133, https://doi.org/10.1016/j.cub.2017.12.051.

Markram, H., Meier, K., Lippert, T., Grillner, S., Frackowiak, R., & Dehaene, S., …Saria, A. (2011). Introducing the Human Brain Project. *Procedia Computer Science, 7*, 39–42, https://doi.org/10.1016/j.procs.2011.12.015.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences, USA, 118*(8), e2011417118, https://doi.org/10.1073/pnas.2011417118.

Mély, D. A., Linsley, D., & Serre, T. (2018). Complementary surrounds explain diverse contextual phenomena across visual modalities. *Psychological Review, 125*(5), 769–784, https://doi.org/10.1037/rev0000109.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. Retrieved from https://distill.pub/2017/feature-visualization/.

Pelli, D. G., Palomares, M., & Majaj, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision, 4*(12):12, 1136–1169, https://doi.org/10.1167/4.12.12.

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11*(10), 1129–1135.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., & Agarwal, S., …Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv*, arXiv:2103.00020v1 [cs.CV].

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.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., & Christensen, A., …Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience, 22*(11),

1761–1770, https://doi.org/10.1038/s41593-019-0520-2.

Rosenholtz, R. (2017). Capacity limits and how the visual system copes with them. *Electronic Imaging, 2017*(14), 8–23, https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-111.

Sayim, B., Westheimer, G., & Herzog, M. H. (2008). Contrast polarity, chromaticity, and stereoscopic depth modulate contextual interactions in Vernier acuity. *Journal of Vision, 8*(8):12, 1–9, https://doi.org/10.1167/8.8.12.

Saxe, A., Nelli, S., & Summerfield, C. (2020). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience, 22*(1), 55–67.

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*, https://doi.org/10.1101/407007.

Schrimpf, M., Kubilius, J., Lee, M. J., Ratan Murty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative Benchmarking to advance neurally mechanistic models of human intelligence. *Neuron, 108*(3), 413–423, https://doi.org/10.1016/j.neuron.2020.07.040.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *ArXiv*. arXiv:1409.1556v6 [cs.CV].

Spang, K., Grimsen, C., Herzog, M. H., & Fahle, M. (2010). Orientation specificity of learning Vernier discriminations. *Vision Research, 50*(4), 479–485.

Storrs, K. R., Khaligh-Razavi, S. M., & Kriegeskorte, N. (2020a). Noise ceiling on the crossvalidated performance of reweighted models of representational dissimilarity: Addendum to Khaligh-Razavi & Kriegeskorte (2014). *BioRxiv*, https://doi.org/10.1101/2020.03.23.003046.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020b). Diverse deep neural networks all predict human IT well, after training and fitting. *BioRxiv*, https://doi.org/10.1101/2020.05.07.082743.

Su, J., Vargas, D. V., & Kouichi, S. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation, 23*(5), 828–841, https://doi.org/10.1109/TEVC.2019.2890858.

Sun, E. D., & Dekel, R. (2019). ImageNet-trained deep neural network exhibits illusion-like response to the Scintillating Grid. *arXiv*, arXiv:1907.09019v2 [cs.CV].

Tao, X., Gao, H., Shen, X., Wang, J., & Jia, J. (2018). Scale-recurrent network for deep image deblurring. *arXiv*, arXiv:1802.01770v1 [cs.CV].

Todorović, D. (2018). In defence of illusions: A reply to Braddick (2018). *Perception, 47*(9), 905–908, https://doi.org/10.1177/0301006618787613.

Todorović, D. (2020). What are visual illusions? *Perception, 49*(11), 1128–1199, https://doi.org/10.1177/0301006620962279.

Toet, A., & Levi, D. M. (1992). The two-dimensional shape of spatial interaction zones in the parafovea. *Vision Research, 32*(7), 1349–1357.

Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., & Unterthiner, T., …Dosovitskiy, A. (2021). MLP-Mixer: An all-MLP architecture for vision. *arXiv*, arXiv:2105.01601v4 [cs.CV].

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology, 8*, 142, https://doi.org/10.3389/fpsyg.2017.00142.

VanRullen, R., & Thorpe, S. J. (2002). Surfing a spike wave down the ventral stream. *Vision Research, 42*(23), 2593–2615, https://doi.org/10.1016/s0042-6989(02)00298-5.

Volokitin, A., Roig, G., & Poggio, T. (2017). Do deep neural networks suffer from crowding? *arXiv*, arXiv:1706.08616v1 [cs.CV].

Ward, E. J. (2019). Exploring perceptual illusions in deep neural networks. *bioRxiv*, https://doi.org/10.1101/687905.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology, 9*, 345, https://doi.org/10.3389/fpsyg.2018.00345.

Xu, H., Ma, Y., Liu, H., Deb, D., Liu, H., Tang, J., … Jain, A. K. (2019). Adversarial attacks and defenses in images, graphs and text: A review. *arXiv*, arXiv:1909.08072v2 [cs.LG].

Xu, Y., & Vaziri-Pashkam, M. (2021). Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature Communications, 12*, 2065, https://doi.org/10.1038/s41467-021-22244-7.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, USA, 111*(23), 8619–8624, https://doi.org/10.1073/pnas.1403112111.

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep DNN for image denoising. *IEEE Transactions on Image Processing, 26*(7), 3142–3155, https://doi.org/10.1109/TIP.2017.2662206.