# Do DNNs explain the visual system?
# Guidelines for a better debate about explanation

**Abstract:** Deep neural networks (DNNs) achieve impressive results in computer vision, translation, and text generation. They are now offered as predictively powerful *models* of neural systems like the ventral visual system. This raises a question that has sparked a debate in the cognitive sciences: if these models predict the neural activity of a system, do they explain how this system works? To help researchers tackle this question, we propose five guidelines: (1) define 'explanation,' (2) specify what about the system the model explains, (3) specify what about the model does the explaining, (4) specify how much explanatory information the model contains, and (5) clarify how much information must be intelligible, and to whom, to explain. We argue that most disagreement about whether DNNs explain divides along these guidelines. We unpack and explicate these guidelines, highlighting why we must consider them whenever we ask whether a model explains something.

*The summary presented here represents an interdisciplinary collaboration stemming from a 2024 workshop at ENS Paris on "Concepts for Understanding Brain Organization."*

**Authors:**

David Colaço
Munich Center for Mathematical Philosophy
LMU Munich
*Corresponding Author*
david.colaco@lmu.de

Russell A. Poldrack
Department of Psychology
Stanford University

Aliya Rumana
Center for Philosophy of Science
University of Pittsburgh

Jordan Theriault
Departments of Psychology and Biology
Northeastern University
Research Affiliate
Department of Radiology
Massachusetts General Hospital

Daniel Burnston
Philosophy Department
Tulane University
Tulane Brain Institute

Sofie Valk
Research group leader
Max Planck Institute for Human Cognitive and Brain Sciences
Research Center Juelich, INM-7, Brain and Behavior
Heinrich Heine University Duesseldorf, Institute for Systems Neuroscience

Philipp Haueis
Institute of Philosophy
Leibniz University Hannover

Daniel Margulies
Centre national de la recherche scientifique
Integrative Neuroscience & Cognition Center
University of Paris

Carl F. Craver
Washington University in St. Louis

# Introduction

Deep neural networks (DNNs) – networks that comprise artificial neurons, arranged in an architecture including multiple hidden layers, whose weighted connections result from training on datasets according to a learning rule – have achieved impressive results in fields like computer vision, translation, and text generation. They are also increasingly offered as *models* of biological systems. For example, the hierarchical structure of DNNs putatively reflects the ventral visual system (VVS), with layers of image processing corresponding to regions in the visual stream. Some modeling successes include the ability to recapitulate receptive field properties of visual neurons in DNN layers, and, strikingly, to *predict* novel visual neuron behaviors. Bashivan, Kar, & DiCarlo (2019), for instance, show that, from a DNN, one can predict neural responses to untrained stimuli much better than traditional models in visual psychophysics.

We can grant that these models successfully predict neural data, but do these models *explain* how this system works? Explanation and prediction are distinct scientific projects, even if they are often pursued in tandem, and even if explanatory models also can be used to make predictions. If we want to tackle this question, we must take a step back and address *what we think scientific explanation is*. This highlights that the debate falls into the domain of philosophy as much as it is in the domains of neuroscience and artificial intelligence.

This topic is our focus in this piece. Unlike previous work, such as an article (Cichy & Kaiser, 2019) that considered the value of DNN models and urged researchers to distinguish between their explanatory, predictive, and exploratory values, we address the question: supposing that these models are predictively successful, how do we determine whether they explain how something like the visual system works? In addressing this question, we reveal the complications hidden within debates over answers to it.

For example, some scientists and philosophers claim that some DNN models explain the workings of the VVS. One part of this claim is that at least some of the structures and organization of the model accurately represent the system itself. Those who think DNNs explain can admit that they are not *complete* explanations. They can allow that we do not have to understand every detail of a model. Rather, they posit that we should understand these DNN-based models of the visual system as "runnable abstractions" (Cao & Yamins, 2024), as "minimal models" (Sullivan, 2022), as instances of "compressibility" (Richards et al., 2019), or in terms of a "high-level code" (Lillicrap & Kording, 2019) that underlies its workings. Scholars in this camp can also claim that the DNN models are *intelligible* (Celeghin et al., 2023; Saxe, Nelli, & Summerfield, 2021), even if we do not understand them in their entirety.

On the other hand, some scholars argue that these DNN models do not explain how the visual system works. Rather, they provide purely *predictive* knowledge about, for example, which category judgments it will produce or how neurons within it will respond to stimuli

(Hasson, Nastase, & Goldstein, 2020; Yarkoni & Westfall, 2017). This perspective is often conjoined with the notion that DNNs are *unintelligible*: we do not know how they do what they do, except that their behavior somehow emerges through the interaction of their components. Consequently, we are limited when mapping the models' structure to the brain (Zhou & Danks, 2020), or worse, we try to explain something we do not understand in terms of something else that we do not understand. Yet, intelligibility might not exhaust the value of DNN models, as some scholars hold that these models provide information that does not involve them being intelligible, such as which parameter choices relate to network performance (Chirimuuta, 2024; Yarkoni & Westfall, 2017).

Further, some individuals (present company included) want the best of both worlds. For instance, Kriegeskorte says, "We must … strive to understand … how exactly the network transforms representations across the multiple stages of a deep hierarchy" but also that "we should be prepared to deal with mechanisms that elude a concise mathematical description and an intuitive understanding" (2015, 438). This example shows that there is disagreement and ambivalence over the explanatory value of DNNs.

Our aim is not to *resolve* these issues in this piece. Instead, it is to clarify what is at stake, to localize genuine points of disagreement, and to help the community avoid chasing red herrings. We think that the flat question "Do DNNs explain the ventral visual system?" is ill-posed, and that philosophy of science can help us to ask better questions and avoid confusions. We offer five guidelines for thinking about the explanatory status of DNN models. These guidelines are important for us to keep in mind whenever we want to assess whether a model explains something.

## Guideline 1: Define 'explanation.'

When we ask whether DNNs explain, we first must have some idea of what explanation is. Correspondingly, we must have some idea of when a model explains. Different accounts of the nature of explanation are available, and disagreements about whether models explain may turn on deciding what explanation is.

Many, though not necessarily all, explanations in neuroscience are causal-mechanistic. They reveal the *causal structure* of the system (Salmon, 1984; Craver, 2007). They show how things come about, how things work, or how things fit into the working of a higher-level system. For example, an explanation of the action potential describes occurrences in the neuron: the flow of current through ion channels in the plasma membrane. This activity contributes to the behavior of the cell populations in which the given cell operates.

Some scholars argue that explanations do not *have* to detail causal structure but can describe topological or dynamic constraints to which a system is subject (Lange, 2016; Ross, 2023). Alternatively, some scholars argue that explanations enable us to situate what we want to explain in a universality class (Chirimuuta 2012; Batterman & Rice 2014). These are

different, though not necessarily mutually exclusive, positions on what it takes for a model to explain.

We also can ask what *cognitive* requirements explanations must satisfy. For some scholars, explanation is tied to understanding: a so-called "toy model" that gives us insight into the organizing principles of a causal structure thereby explains it, even if the model is inaccurate (Beer et al., 2024). Other scholars argue that understanding is dissociable from explaining. On this view, a model can be explanatory even if this model is only minimally understood (Craver & Kaplan, 2020).

Disagreements about whether DNN models explain may turn on disagreements about what one requires of explanations more generally. If one requires a highly accurate description of the causal structure of the visual system for explanation, one might take the lack of detailed match between DNNs and visual system structure to show that DNNs are not explanatory (Bowers et al., 2022). However, if one requires only a minimal degree of accuracy, so long as the model increases our understanding, one is more likely to view the models as explanatory (Bechtel 2016). We might, for instance, think that an explanation can give us insight into the image recognition mechanism without revealing much of the causal structure, and we might conclude that DNN models do explain the ventral visual system (Cao & Yamins, 2024). Differing positions about whether DNNs explain thus might amount to differing positions on the roles of accuracy, causal detail, and understanding when we assess whether a model explains something.

**Lesson:** Asking whether DNNs explain requires clarifying *what one takes an explanation to be*. This requires clarifying both what features of the system must be captured in a model and what kind of cognitive requirement (if any) explanation must satisfy.

## Guideline 2: Specify what you are trying to explain.

Explaining requires specifying *what phenomenon* one wants to explain. If a scientist wants to explain the visual system, we should probe what about the system they aim to explain. Is it phototransduction? Neuronal response profiles? Object recognition? Color vision? These call for different explanations, focusing on different features of the system. A DNN model might explain one of these phenomena without addressing the others.

Two scientists with different explanatory questions in mind might disagree about whether a DNN model explains, even if they are both right for their own question. So we should ask not "Do DNNs explain?" but rather "What (if anything) does *this* DNN tell us that is relevant to *this phenomenon*?" Although scientists sometimes speak in general terms about *the* explanation for vision, for example, this is shorthand for myriad explanations of different features of vision.

Complicating matters, scientists' ideas about phenomena are not *static*. They often revise, refine, or reject them as they learn more about the system (Colaço, 2020), changing the explanatory questions they ask. If the question changes, different models, or changes to existing models, are often necessary. We see this in the DNN literature. DNNs are used to model instances of perceptual learning (Wenliang & Seitz, 2018), to account for the intrinsic memorability of images (Needell & Bainbridge, 2022), and to account for perceived object similarity as opposed to just object classification (Devereux, Clarke, & Tyler, 2018; Jozwik et al. 2023). The first two cases involve distinct architectural and training regime assumptions for existing DNN models. The similarity case, by contrast, involves both making specific modeling assumptions about the DNN *and* combining it with non-DNN "visuo-semantic" models. In terms of neural data, accounting for neural dynamics in addition to response properties requires modeling recurrent connections that other DNNs leave out (Kietzmann et al., 2019). These examples show that the phenomena researchers aim to explain can be both different and changing.

A substantive objection to a particular DNN model would be a criticism of its ability to answer the explanatory question for which it was intended. Some claim, for instance, that DNNs do not actually capture categorization behavior (Rajalingham et al., 2018), and the famous case of texture bias putatively shows that DNNs do not explain the way *humans* categorize (Bowers et al., 2022). These are objections about specific models and specific explanatory questions. They are not objections about the status of DNNs as explanatory models more generally.

**Lesson:** Before we ask whether DNNs explain, we should be clear about what we want to explain and whether we agree on the explanatory question that we want to answer. While there is a range of data that DNNs allegedly *fail* to capture (Bowers et al., 2022; Yuille & Liu, 2020), this alone does not disqualify DNNs as explanations more generally.

## Guideline 3: Consider how modeling and explanation relate.

Suppose we agree on what explanation is and what we want to explain. Now, we must turn to scientific *models*, which are representations. They are *abstractions*: they describe some features of the system and leave out others. They are also *idealizations*: they include features and assumptions that are false or that have no corresponding referent in the system.

By consequence, one must clarify *which* properties of the model one attributes to the system. If a model of edge-detection works by computing a difference of Gaussians, for instance, it is an extra step to say that cells in the lateral geniculate nucleus also detect edges by computing this difference. Moreover, one must choose the degree of abstraction of one's attributions from the model to the world. One can posit that LGN cells compute a difference of Gaussians without positing that they precisely recapitulate the shape of the

function in the model. The fit between a model and the world is always imperfect. The modeler uses abstract or idealized representations to make certain aspects of the system salient.

The commitments of the model might not be explicit. Consider commitments about a model's architecture. No one contends that the number of units in a hierarchical layer in their model is the *same* as the number of neurons in the relevant part of the visual system that they aim to model. "Number of units" is not something attributed to the world from the model. Other times, these commitments are explicit, such as when a modeler says that hierarchical DNN levels *do* correspond to areas along the primate ventral visual stream. Sometimes, modelers attribute some features of a model to the system while not attributing others. For instance, the *hierarchical* structure of a model might be attributed to the system, without the *purely feed-forward* connectivity or the learning rule being attributed along with it (Celeghin et al., 2023). Similarly, one might attribute informational transforms between layers in a network to the system but *not* attribute to it the network's training regime, which may well be biologically implausible.

A substantive objection to make to a particular DNN model would be that it does not accurately capture the features that a modeler attributes to the system. For instance, Sexton & Love (2022) mapped physiological activity in given brain areas to model activation in a DNN, thereby investigating how this activity affected categorization performance. They showed that physiological activity from even earlier areas recapitulated model performance only when mapped to activity in a high layer of the network. This is an objection to claims that the response profile of successive DNN layers should be attributed to the visual system.

**Lesson:** *All* models are abstract and idealized, so it is not an inherent problem that a DNN model is abstract and idealized. What matters is whether the model is accurate with regard to the features that are attributed to the system.

## Guideline 4: Recognize that models are partial.

Suppose we accept that a DNN model explains a phenomenon of interest. This does not mean that we must accept that this one model completely answers one explanatory question, let alone all of them. Models are inherently *partial*, meaning that *single models rarely provide complete explanations* (Hochstein 2017). For example, endeavors to explain the action potential likely will include models of ion-channel structure, models based on the Nernst and Goldman equations for membrane potential and conductance, oscillator models that map changes in membrane permeability to changes in potential over time, pathway descriptions of second-messenger cascades that modify channel permeability, and others. Likewise, whether discussing DNNs or any other model in neuroscience, one model need not be intrinsically better than another. Rather, they may pick out different things about what we want to explain, which we might combine when producing an explanation (Hochstein 2016). Each model contributes to, but none fully constitutes, the explanation.

The answer to the question "does this model explain X" is typically not all-or-nothing. We are better off asking *what* or *how much* explanatory information a model conveys. A tendency to eat ultra-processed foods, for example, might provide explanatory information without fully explaining one's body weight. DNNs are no different. A model might capture a learning process without accurately describing neural data. By contrast, a model might capture the sequential organization of processing stages but not represent the processing at each stage. Each conveys explanatory information, but none are complete explanations. An aspect of a DNN model, such as its hierarchical structure, might convey explanatory information, while other components, such as the learning algorithm, might not. Similarly, a training regime might partially account for how visual learning works, but a fuller account might incorporate continuous learning (Saxe, Nelli & Summerfield, 2021) or inductive biases (Richards et al., 2019).

**Lesson:** One should think of a model as an *explanatory resource*. Rather than asking whether a model explains full stop, we should ask how much explanatory information this model conveys about the phenomenon of interest.

## Guideline 5: Recognize that intelligibility is neither fixed nor all-or-nothing.

Even if we agree about explanation so far, a common complaint is that DNN models are not explanations because they are difficult for us to understand. In other words, the complaint is that these models have low *intelligibility* to us. This complaint seems to presuppose that explanation has something to do with understanding, which *Guideline 1* shows is itself controversial. Even if we take for granted this connection between explanation and understanding, though, we must be clear about exactly *what* must be intelligible.

Intelligibility can apply to either a model or the mapping relationship from the model to what it putatively represents (Sullivan, 2022). Both concerns are relevant for DNN models of the visual system. On one hand, we might doubt that DNNs can explain the visual system because their own functional organization is difficult to understand. On the other hand, it is also difficult to interpret the mappings from activation vectors in the DNN model to firing rates in the ventral visual stream, which are identified with linear regression and then used to predict neural data. Cao & Yamins (2024), for instance, suggest that this mapping from the model to the target is best understood as an abstract equivalence between operations in the DNN model and activities in the ventral visual stream, but exactly how this works is far from obvious.

If we doubt that DNN models are explanatory due to model intelligibility, we may be encouraged by progress in "transparent AI" (Räuker et al., 2023), which aims to increase our understanding of DNNs. But, if we doubt that DNN models are explanatory due to mapping

intelligibility, then we may be more encouraged by progress in the study of generalization in deep learning (Zhang et al., 2017; Kawaguchi et al., 2023), which aims to increase our understanding of how features in the world drive the predictive successes of DNN models.

At the same time, assessments of a model's intelligibility can change over time and with changes in historical and intellectual contexts. One generation's unintelligible posit can become the bedrock of another's science, as it has time and again. Scientists well-versed in training and working with DNNs will likely find their inner workings less mysterious than researchers who are looking in from the outside. Whether a given theoretical posit counts as "intelligible" to a group of scientists is context-specific, can change with time, and can be learned as scientists become increasingly familiar with "new" ideas.

**Lesson:** If explanation requires intelligibility, we must clarify whether this intelligibility has to do with the model itself or its mapping onto the phenomena it represents. We should also clarify *what degree* of intelligibility might be satisfactory for explanation, recognizing that the answer will depend on context.

## Conclusion

We do not intend to *answer* the question of whether DNNs explain the workings of the ventral visual system in this paper. Rather, our aim has been to show that flatly asking questions like "do DNNs explain?" is insufficient. This simple-sounding question hides many disagreements that we hope we have made more salient with our guidelines. We hope that putting some constraints on the question, its meaning, and what will be needed out of an answer to it will help to guide more fruitful discussions.

The undeniable predictive successes of DNNs alone should not lead us to assume that these models explain, let alone to assume that they explain everything we want to explain. History corroborates the idea that there are no silver bullets when it comes to explaining biological systems like the brain (Mitchell & Gronenborn, 2017). DNNs are no exception, and we should not buy uncritically into the explanatory hype around them. At the same time, these models should not be held to the unreasonable standard of providing a complete, accurate, or fully intelligible model of what we aim to explain when we aim to explain things about neural systems. The fact that DNNs are not complete explanations does not impugn their potential as contributions to our explanations of how the brain works. A reflective look at what explanation requires, we hope, will give us a more nuanced ability to assess their explanatory potential. Further, it will help point us towards a more productive discussion.

# References


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

Batterman R.W., & Rice, C.C. (2014). Minimal Model Explanations. Philosophy of Science, 81(3): 349-376. doi:10.1086/676677

Bechtel, W. (2016). Using computational models to discover and understand mechanisms. *Studies in History and Philosophy of Science Part A*, *56*, 113-121. https://doi.org/10.1016/j.shpsa.2015.10.004

Beer, R. D., Barwich, A. S., & Severino, G. J. (2024). Milking a spherical cow: Toy models in neuroscience. *European Journal of Neuroscience*, *60*(10), 6359-6374. https://doi.org/10.1111/ejn.16529

Bowers, J. S., Malhotra, G., Dujmovic, M., Llera Montero, M., Tsvetkov, C., Biscione, V., . . . Blything, R. (2022). Deep problems with neural network models of human vision. *Behav Brain Sci, 46*, e385. doi:10.1017/S0140525X22002813

Cao, R., & Yamins, D. (2024). Explanatory models in neuroscience, Part 1: Taking mechanistic abstraction seriously. *Cognitive Systems Research*, 101244.

Celeghin, A., Borriero, A., Orsenigo, D., Diano, M., Méndez Guerrero, C. A., Perotti, A., ... & Tamietto, M. (2023). Convolutional neural networks for vision neuroscience: significance, developments, and outstanding issues. *Frontiers in Computational Neuroscience*, *17*, 1153572. https://doi.org/10.3389/fncom.2023.1153572

Chirimuuta, M. (2024). *The Brain Abstracted: Simplification in the history and philosophy of neuroscience*: MIT 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

Craver, C. F. (2007). Explaining the Brain. Oxford, GB: Oxford University Press.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy015

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Sci Rep, 8*(1), 10636. doi:10.1038/s41598-018-28865-1

Hochstein, E. (2016). One mechanism, many models: a distributed theory of mechanistic explanation. *Synthese* 193, 1387–1407. https://doi.org/10.1007/s11229-015-0844-8

Hochstein, E. (2017). Why one model is never enough: a defense of explanatory holism. *Biology & Philosophy*, *32*, 1105-1125. https://doi.org/10.1007/s10539-017-9595-x

Jozwik, K. M., Kietzmann, T. C., Cichy, R. M., Kriegeskorte, N., & Mur, M. (2023). Deep Neural Networks and Visuo-Semantic Models Explain Complementary Components of Human Ventral-Stream Representational Dynamics. *J Neurosci, 43*(10), 1731-1741. doi:10.1523/JNEUROSCI.1424-22.2022

Kawaguchi, K., Deng, Z., Ji, X., & Huang, J. (2023). How does information bottleneck help deep learning?. In *International Conference on Machine Learning* (pp. 16049-16096). PMLR.

Kietzmann, T. C., Spoerer, C. J., Sorensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proc Natl Acad Sci U S A, 116*(43), 21854-21863. doi:10.1073/pnas.1905544116

Lange, M. (2016). *Because without cause: Non-casual explanations in science and mathematics*. Oxford University Press.

Lillicrap, T. P., & Kording, K. P. (2019). What does it mean to understand a neural network? *arXiv preprint arXiv:1907.06374*.

Mitchell, S. D., & Gronenborn, A. M. (2017). After fifty years, why are protein X-ray crystallographers still in business?. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axv051

Needell, C. D., & Bainbridge, W. A. (2022). Embracing New Techniques in Deep Learning for Estimating Image Memorability. *Computational Brain & Behavior, 5*(2), 168-184. doi:10.1007/s42113-022-00126-5

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. *J Neurosci, 38*(33), 7255-7269. doi:10.1523/JNEUROSCI.0388-18.2018

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023, February). Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023*

*ieee conference on secure and trustworthy machine learning (satml)* (pp. 464-483). IEEE. doi:10.1109/SaTML54575.2023.00039.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., . . . Ganguli, S. (2019). A deep learning framework for neuroscience. *Nature neuroscience, 22*(11), 1761-1770.

Ross, L.N. (2023). The explanatory nature of constraints: Law-based, mathematical, and causal.*Synthese* 202, 56. https://doi.org/10.1007/s11229-023-04281-5

Salmon, Wesley C. (1984). Scientific Explanation and the Causal Structure of the World. Princeton University Press.

Saxe, A., Nelli, S., & Summerfield, C. (2021). If deep learning is the answer, what is the question?. *Nature Reviews Neuroscience*, *22*(1), 55-67. https://doi.org/10.1038/s41583-020-00395-8

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science advances, 8*(28), eabm2219.

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz035

Wenliang, L. K., & Seitz, A. R. (2018). Deep Neural Networks for Modeling Visual Perceptual Learning. *J Neurosci, 38*(27), 6028-6044. doi:10.1523/JNEUROSCI.1620-17.2018

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci, 12*(6), 1100-1122. doi:10.1177/1745691617693393

Yuille, A. L., & Liu, C. (2020). Deep Nets: What have They Ever Done for Vision? *International Journal of Computer Vision, 129*(3), 781-802. doi:10.1007/s11263-020-01405-z

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). *Understanding deep learning requires rethinking generalization* (arXiv:1611.03530). arXiv. https://doi.org/10.48550/arXiv.1611.03530

Zhou, Y., & Danks, D. (2020). Different" intelligibility" for different folks. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 194-199).