

# Interpretability of artificial neural network models in artificial intelligence versus neuroscience

Kohitij Kar, Simon Kornblith & Evelina Fedorenko



The notion of ‘interpretability’ of artificial neural networks (ANNs) is of growing importance in neuroscience and artificial intelligence (AI). But interpretability means different things to neuroscientists as opposed to AI researchers. In this article, we discuss the potential synergies and tensions between these two communities in interpreting ANNs.

In neuroscience, interpretability often implies an alignment to brain constructs. Conversely, in AI, the emphasis is on making the models’ decision-making process more transparent and explicable to a human interpreter, as needed for understanding social and legal consequences. We argue that attempts to make ANNs more interpretable to neuroscientists should not be conflated with ongoing efforts in explainable AI. However, both AI researchers and neuroscientists can leverage the synergy between neuroscience and AI in working towards interpretable ANN models. In particular, the degree of alignment between ANNs and primate brains and behaviour can serve as a useful benchmark for explainable AI.

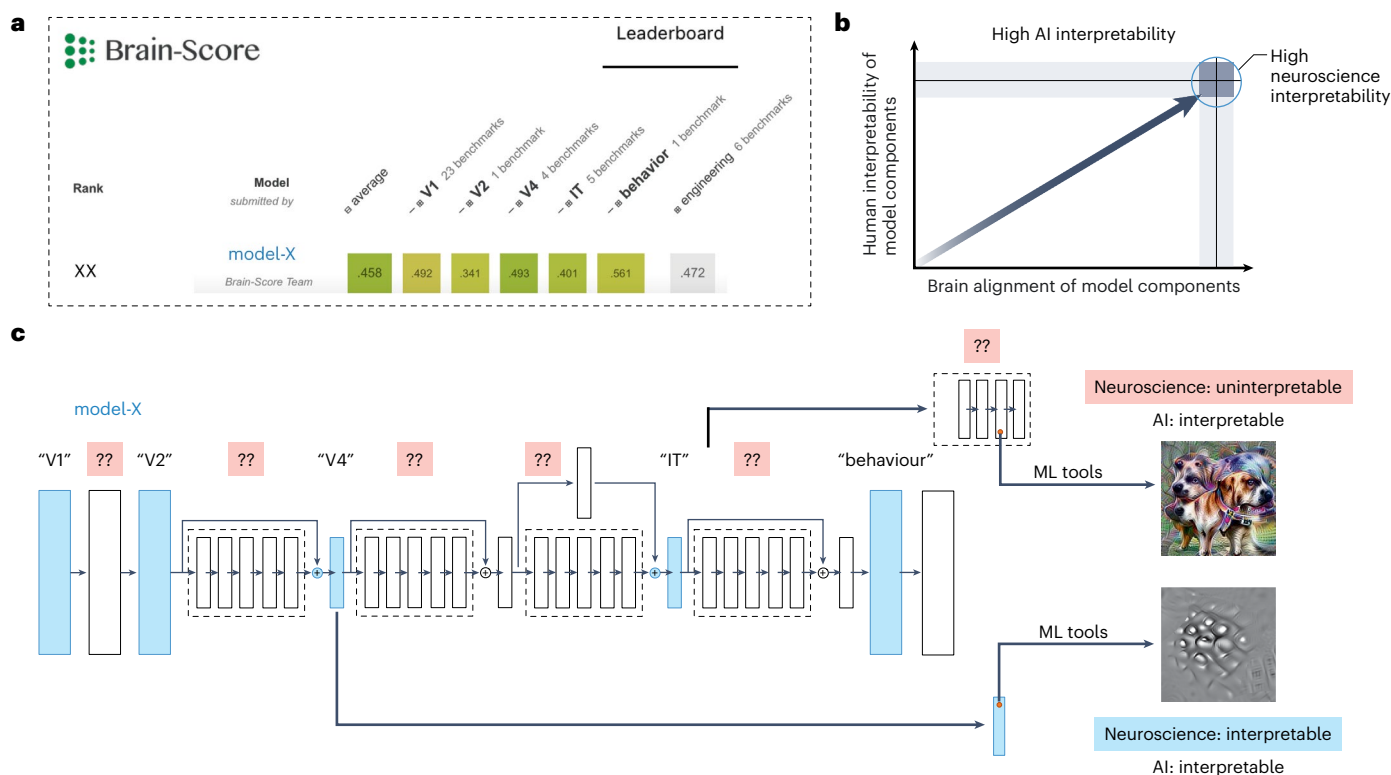
Computationally explicit hypotheses of brain function derived from machine learning (ML)-based models have recently revolutionized neuroscience<sup>1–4</sup>. Despite the unprecedented ability of these ANNs to capture responses in biological neural networks (brains) (Fig. 1a; see<sup>5</sup> for a comprehensive review), and our full access to all internal model components (unlike in the brain), ANNs are often referred to as ‘black boxes’ with limited interpretability. Interpretability, however, is a multi-faceted construct that is used differently across fields. In particular, interpretability and explainability efforts in AI focus on understanding how different model components contribute to its output. By contrast, the neuroscientific interpretability of ANNs requires explicit alignment between model components and neuroscientific constructs like recurrence<sup>6</sup> or top-down feedback<sup>7</sup>. Given the widespread calls to improve the interpretability of AI systems<sup>8</sup>, we here highlight these different notions of interpretability and argue that the neuroscientific interpretability of ANNs can be pursued in parallel with, but independently from, the ongoing efforts in AI. Certain ML techniques, such as DeepDream<sup>9</sup>, can be leveraged in both fields to ask what stimulus optimally activates the specific model features (feature visualization by optimization), or how different features of the input contribute to the model’s output (feature attribution). However, without appropriate brain alignment, certain features will remain uninterpretable to neuroscientists (for instance, the non-blue segments of the model in Fig. 1c).

## A conceptual framework

Like interpreters of human languages, scientists seek a high-fidelity mapping between two ‘languages’: the ‘language’ of scientific measurement, and the ‘language’ of scientific hypotheses (models). The language of measurement consists of numerical descriptions of a sample from the phenomenon we seek to understand. For instance, systems neuroscientists that are interested in visual processing could measure and summarize neural spiking activity from individual visuocortical neurons or obtain behavioural measurements on specific visual tasks. The language of scientific hypotheses consists of conceptual abstractions that aim to explain, predict and control the phenomenon of interest (for example, the pattern of firing rates across neurons predicting the category of visual objects<sup>10</sup> in parts of the ventral visual stream). To claim that a specific model, or parts of it, is uninterpretable to a neuroscientist then means that certain components or features of the model do not map onto any empirically verifiable neuroscientific construct. To our knowledge, all current ANN models of primate vision<sup>3</sup> contain features that have not been explicitly mapped onto neuroscientific constructs, which limits their interpretability.

For example, consider the schematic of an ANN model in Fig. 1c, where certain specific components (in blue) have brain-mapped labels (for example, areas V1, V2 and V4). This mapping is typically achieved by comparing available brain data with representations from different model layers and identifying the best match (for details, see<sup>5</sup>). However, most ANN models, such as of vision, language and audition, contain layers that have no clear mapping to responses from any particular brain region or to any particular computation known to characterize information processing in the relevant domain. It is therefore unclear how to treat these unmapped portions of the model. Do they correspond to neuroscientific constructs that cannot be resolved owing to insufficient available data (for example, individual cortical layers or subregions of larger anatomical regions) or to neuroscientific constructs that have yet to be discovered? Or are they model-specific idiosyncrasies that do not correspond to the brain in any way?

These questions potentially render ANNs not fully interpretable to neuroscientists, who have no way to engage with these unmapped components when designing an experiment or making model-based inferences. A model that only contains brain-mapped components could be more neuroscientifically interpretable – that is, such a model can be used to make brain-related predictions and evaluated or falsified by neural data. We thus advocate consideration of the neuroscientific interpretability of ANNs in addition to current neural and behavioural benchmarks available in Brain-Score<sup>5</sup> and other similar integrative benchmarking platforms<sup>11</sup>. Although further work is necessary to determine how to precisely quantify neuroscientific interpretability, a simple metric is the number of non-brain-mapped components, such that models with fewer such components should be preferred. Taking



**Fig. 1 | Interpretability of models for AI and neuroscience.** **a**, Schematic depiction of the Brain-Score platform<sup>5</sup> website as an example of an integrative benchmarking platform for models of the primate brain<sup>10</sup>. Columns contain scores for different brain benchmarks (for example, V1, V2, V4 and IT predictivity), and each row corresponds to a specific model (example shown here as a dummy, model-X). **b**, How brain alignment of models and human interpretability of model decision-making relates to high AI and neuroscientific interpretability of models. For instance, higher levels of human interpretability of feature attribution (how different features of the input contribute to the model's output) might make ANNs highly interpretable for AI (refer to the top left part of the plot), but poor brain

alignment will lead to low interpretability for neuroscientists. On the other hand, higher brain alignment and a high level of human interpretability will lead to more interpretable models for both AI and the neurosciences (top right). **c**, Schematic of 'model-X', which could be replaced by any current brain-like ANN model. We show an example of how only some parts (in blue) of the model can be interpretable to a neuroscientist (for example with the deep image synthesis tool from ref.<sup>16</sup>) as they can directly map to individual brain areas. As we also show, ML tools such as DeepDream<sup>9</sup> can interpret the entire model for AI (for example, both blue and white boxes). The 'dog on leash' feature visualization image is reproduced from Olah et al.<sup>24</sup>.

seriously the degree to which a model is neuroscientifically interpretable might also begin to address common issues with integrative benchmarking platforms like Brain-Score. For instance, models with more versus less biologically plausible architectures (where biologically plausible means consistency with primate brain anatomy) sometimes perform similarly on existing benchmarks<sup>12</sup>. Taking their interpretability into account can help rank these high-performing models.

We further propose that neuroscientific interpretability is itself a relative term. In particular, the extent to which a model needs to be accessible to a human experimenter, and aligned with neuroscientific constructs, depends on the model's intended use. For instance, a model that is expected to predict the responses in the fMRI-based voxels in specific subregions of the human brain need not map onto the lower-level components of the brain like the neurons. It should, however, have explicit mapping onto all accessible experimental components like the ability to engage with the exact stimulus and the ability to perform the behavioural task. Therefore, a model that is interpretable for one set of experiments may not be interpretable for another. This task-dependent interpretability, however, should

not necessarily discourage models from comprising finer-grained details of the brain but will require the modeler to minimally commit to an explicit mapping between model features and specific relevant experimental variables of interest.

### Leveraging the synergies

One avenue to further explore is using neuroscientific interpretability as a benchmark for the goodness of ANN explainability. There has been growing interest and legislation of AI research across many leading nations to promote and achieve explainable AI<sup>6,13</sup>. However, no ground truth exists for what constitutes a good explanation. Indeed, one of the important challenges for the current AI 'explainability' results is that different methods to interpret the functional role of model features lead to different results and inferences<sup>14</sup>, and it is unclear how to evaluate explanation quality. We propose that one way to validate the 'goodness' of explanations is to measure their match with the explanations derived from primate behaviour<sup>15,16</sup> and neural measurements<sup>17,18</sup>.

For example, gradient-based feature attribution methods such as Grad-CAM<sup>19</sup> can be used to identify which pixels of an image are

responsible for identifying a dog in an image. The explanations that these methods produce take the form of 2D saliency maps of the same size as the input image, reflecting the contribution of each pixel to the network's prediction. Similarly, in human visual psychophysics, techniques such as Bubbles<sup>14</sup> are often used to reveal the image regions that are most informative for decision-making. One approach could be to benchmark the pixel-by-pixel agreement in the ANN and primate behaviour-based saliency maps as a measure of the goodness of the AI explanation. Attribution techniques in AI for ANNs also include layer-level (assessing the contribution of individual model layers) or neuron-level (assessing the contribution of an individual ANN neuron unit) attributions. Methods such as deep image synthesis applied to the brain (as demonstrated in<sup>16</sup>, Fig. 1c), at the level of brain areas or at the single neuron level, also allow us to benchmark the alignment of layer and neuron attribution results between ANNs and primates. A high combined score on such benchmarks will place the models at the top right quadrant of Fig. 1b, which makes them highly interpretable for both AI and neuroscience. Appropriate ceiling estimation remains an important factor that will affect these estimates. For instance, the reliability of the explanations that are produced by a specific feature attribution method applied on a specific ANN under different weight initializations could set the ceiling for comparisons with other feature attribution methods and explanations that are derived from human behaviour.

We are not necessarily proposing that the brain is an optimal system that AI should mimic. But given that the errors made by top-performing ML models increasingly resemble those made by humans<sup>20</sup>, it is reasonable to expect some degree of mechanistic alignment. Thus, in the absence of any ground truth in AI explainability, we argue that similarity with the primate brain (a system that is robust, flexible and capable of powerful generalization) might provide valuable guidance. Quantitatively assessing alignment between ML feature attribution measures and factors that are critical for primate decision-making and brain activation could therefore serve as a putative benchmark for AI explainability measures.


Another avenue to pursue is to make use of ANNs for neuroscience beyond brain alignment. Above, we proposed that a lack of alignment between model components and known neuroscientific constructs decreases an ANN's neuroscientific interpretability (even when reliable AI explanations exist for the models' role in a specific behaviour). However, non-brain-mapped model components can also be beneficial for neuroscience. Indeed, we acknowledge a long tradition where neuroscientists have benefitted from drawing abstract analogies between brains and ANNs<sup>3,4,21–23</sup>. In particular, aspects of models that do not map onto known brain mechanisms may be critical for discovering new mechanisms. In other words, neuroscientifically 'uninterpretable' (in the current quantitative sense) ANN components can act as hypotheses to be tested in future neuroscience experiments and thus expand the repertoire of neuroscientific constructs and become interpretable in the long run.

To conclude, we encourage researchers to conceptually separate the objectives of AI and neuroscience while interpreting the parameters and operations of current computational models. However, we also suggest that – although current AI models operate differently from

primate brains in various ways – all else being equal, interpretability methods in AI that provide more primate-brain-aligned model interpretations are likely to be more promising. In turn, components of current models that do not map onto known neuroscientific constructs could inspire new ideas about how biological brains work.

**Kohitij Kar** <sup>1,2,3</sup> , **Simon Kornblith**<sup>4</sup> & **Evelina Fedorenko** <sup>2</sup>

<sup>1</sup>Department of Biology, Centre for Vision Research, York University, Toronto, Ontario, Canada. <sup>2</sup>McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>3</sup>Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>4</sup>Brain Team, Google Research, Toronto, Ontario, Canada.

 e-mail: [kOh1t1j@yorku.ca](mailto:kOh1t1j@yorku.ca)

Published online: 19 December 2022

## References

1. Yamins, D. L. K. et al. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
2. Schrimpf, M. et al. *Proc. Natl Acad. Sci. USA* **118**, e2105646118 (2021).
3. Pospisil, D. A., Pasupathy, A. & Bair, W. *Elife* **7**, e38242 (2018).
4. Bao, P., She, L., McGill, M. & Tsao, D. Y. *Nature* **583**, 103–108 (2020).
5. Schrimpf, M. et al. Preprint at *bioRxiv* <http://biorxiv.org/lookup/doi/10.1101/407007> (2018).
6. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. *Nat. Neurosci.* **22**, 974–983 (2019).
7. Kar, K. & DiCarlo, J. J. *Neuron* **109**, 164–176 (2021).
8. European Parliament. Directorate General for Parliamentary Research Services. *A governance framework for algorithmic accountability and transparency*. (Publications Office, 2019).
9. Mordvintsev, A., Olah, C., & Tyka, M. Inceptionism: Going deeper into neural networks (2015); <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>
10. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. *J. Neurosci.* **35**, 13402–13418 (2015).
11. Willeke, K. F. et al. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2206.08666> (2022).
12. Conwell, C. et al. SVRHM 2021 Workshop (NeurIPS, 2021).
13. Holzinger, A. in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)* 55–66 (IEEE, 2018).
14. Hooker, S., Erhan, D., Kindermans, P.-J. & Kim, B. *Advances in Neural Information Processing Systems* **32** (2019).
15. Gosselin, F. & Schyns, P. G. *Vision Res.* **41**, 2261–2271 (2001).
16. Murray, R. F. *J. Vis.* **11**, 2 (2011).
17. Bashivan, P., Kar, K. & DiCarlo, J. J. *Science* **364**, eaav9436 (2019).
18. Ponce, C. R. et al. *Cell* **177**, 999–1009 (2019).
19. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. In *International conference on computer vision* 618–626 (IEEE, 2017).
20. Geirhos, R. et al. *Advances in Neural Information Processing Systems* **34**, 23885–23899 (2021).
21. Zipser, D. & Andersen, R. A. *Nature* **331**, 679–684 (1988).
22. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. *Nature* **503**, 78–84 (2013).
23. Olshausen, B. A. & Field, D. J. *Nature* **381**, 607–609 (1996).
24. Olah, C., Mordvintsev, A., Schubert, L. Feature Visualization (Distill, 2017); <https://distill.pub/2017/feature-visualization>

## Acknowledgements

The authors would like to thank C. Shain for helpful comments and discussions. E.F. was supported by NIH awards R01-DC016607, R01-DC016950 and U01-NS121471, and by research funds from the McGovern Institute for Brain Research, the Brain and Cognitive Sciences Department and the Simons Center for the Social Brain. K.K. was supported by the Canada Research Chair Program. This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund. K.K. was supported by an unrestricted research fund from Google LLC.

## Competing interests

The authors declare no competing interests.

## Additional information

**Peer review information** *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.