Brittle Brain Encoding: Poor Out-of-Distribution Generalization Shows the Human Brain is neither a Nintendo Entertainment System nor a Four-Layer Convolutional Neural Network

Yann Harel* (yann.harel@umontreal.ca)
Université de Montréal, Département de psychologie, Montréal, Québec, Canada
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, Québec, Canada

François Paugam* (francois.paugam@umontreal.ca)
Université de Montréal, Département d'informatique, Montréal, Québec, Canada
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, Québec, Canada
Mila – Québec Artificial Intelligence Institute, Montréal, Québec, Canada

Marie St-Laurent (stlaurent.marie@criugm.qc.ca)
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, Québec, Canada

Lune Bellec (lune.bellec@umontreal.ca)
Université de Montréal, Département de psychologie, Montréal, Québec, Canada
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, Québec, Canada
Mila – Québec Artificial Intelligence Institute, Montréal, Québec, Canada

Abstract

To explore the correspondence between artificial neural networks and brain function, we tested three models—trained agent, untrained agent, and game RAM—on their capacity to use game data to predict brain activity (fMRI) in humans playing *Super Mario Bros.* All brain encoding models performed similarly within the training distribution (training levels), but none generalized to out-of-distribution (OOD) levels. The OOD performance drop was generally greater than the difference between models. Our results underscore how current brain encoding approaches may overstate brain-model similarity, and highlight the critical importance of evaluating generalization when using brain scores to compare models.

Keywords: brain encoding; reinforcement learning; out-of-distribution generalization; model interpretability;

Introduction

Many brain encoding models use AI-derived latent representations to predict neural activity from complex stimuli. Voxel-wise models have effectively mapped passive perception of natural images, videos, and text onto brain responses (e.g. Caucheteux & King, 2022; Gifford, Bersch, et al., 2025; Gifford, Cichy, et al., 2025). Additionally, deep neural networks trained on behavioral tasks have emerged as powerful feature extractors, improving neural prediction across both passive and active domains (Yamins et al., 2014; Cross et al., 2021; Kemtur, 2024).

A common assumption is that high encoding accuracy reflects convergence between artificial and biological representations (Schrimpf et al., 2020). Some suggest that encoding models approach the "noise ceiling" on certain benchmarks, implying that AI models may process specific stimuli in genuinely brain-like ways (Caucheteux & King, 2022). This view aligns with the "Platonic representation hypothesis" (Huh et al., 2024), which posits that internal representations, shaped by the statistical structure of the world, naturally converge across systems regardless of architecture.

However, high accuracy does not guarantee functional convergence. Models may overfit stimulus-specific details rather than capturing true abstractions (Gifford, Cichy, et al., 2025). A more stringent test is out-of-distribution (OOD) generalization: predicting brain responses to novel contexts unseen during training. OOD evaluation helps determine whether apparent convergence reflects meaningful functional alignment or superficial overfitting (Shirakawa et al., 2024).

We tested OOD generalization in a naturalistic videogame setting using whole-brain fMRI from participants playing *Super Mario Bros.* (Nintendo, 1983). We evaluated three encoding models: (1) a PPO-trained reinforcement learning agent, (2) an untrained network with the same architecture, and (3) the raw emulator RAM state. Brain predictions were compared on training levels (within-distribution) and unseen levels from the same game (OOD).

Of note, we did not expect models to truly converge with the brain. Reinforcement learning agents are notoriously brittle, often failing to generalize to even minor changes in context—unlike humans. Untrained networks encode random features, and a NES RAM encodes everything in the game, both in a rather unbrain-like fashion. Based on the work by (Cross et al., 2021), we expected within-distribution encoding performance to follow the order PPO > Untrained ≥ RAM. Importantly, we also predicted a dramatic drop in performance for all three models during OOD encoding, which would bring quantitative evidence that the brain's functional representations diverge substantially from those of a NES videogame console or a four-layer convolutional neural network.

Methods

Data. We used fMRI and behavioral data from the Courtois Neuromod Project (https://cneuromod.ca). Five participants played 22 levels of Super Mario Bros. during fMRI scanning (total: 84 h; 13–18 h per participant). Functional images were acquired on a 3T Siemens Prisma Fit scanner (TR = 1.49 s; 2 mm isotropic), preprocessed with fMRIPrep (Esteban et al., 2018; v20.2 LTS), and projected onto the MIST atlas (1,095 parcels; Urchs et al., 2019).

Brain encoding experiments. A four-layer convolutional neural network was trained to play 20 levels of Super Mario Bros. using PPO reinforcement learning (Schulman et al., 2017) with rewards based on forward movement and survival. We extracted activations from the third layer (3,872 features). The same procedure was repeated with an untrained network of identical architecture. Encoding models were constructed using ridge regression to predict parcel-wise BOLD responses from these features, using participants' gameplay replays as ANN input. In parallel, we trained a ridge regression model using the game's RAM state (accessed via gym-retro; Nichol et al., 2018), reduced from 13,321 to 3,872 features via random projection.

Model validation. Performance was evaluated on both within-distribution (WD) and out-of-distribution (OOD) tasks. WD scores were computed on held-out replays from the 20 training levels (80/10/10% train/val/test, stratified by level completion). OOD

scores were computed on replays from two novel levels (5–1 and 6–3) unseen during training.

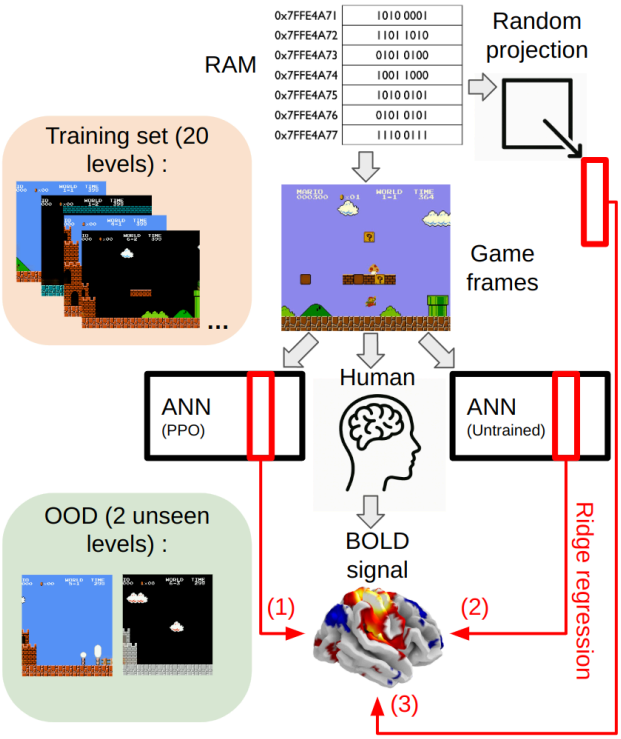

Figure 1: Schematics of the brain encoding pipeline. Human participants played the videogame Super Mario Bros. on a console emulator. Three brain encoding experiments were then performed using activations of an intermediary layer from (1) a reinforcement learning agent trained with PPO and (2) an untrained agent of identical architecture, or (3) using the game memory states (RAM) after random projection to reduce dimensionality.

Results

Brain encoding scores reflect the proportion of variance explained in each brain parcel. The **WD encoding score maps were qualitatively similar** between the three models (Fig. 2A, average across subjects). Encoding performance peaked in the ventral and dorsal visual networks as well as the dorsal attentional network (with R2 about 0.3). WD differences between models were small but significant, with PPO > Untrained > RAM for almost all subjects. **The OOD encoding score maps were dramatically lower than WD** (Fig. 2B, average across subjects), although the topography of the maps remained consistent with WD, peaking in the same networks. Across most subjects and models, the effect of domain shift (i.e., the drop from WD to OOD encoding performance) exceeded any differences observed between model types under either condition (Fig. 2C). This result indicates that **the distribution shift had a much stronger impact on encoding performance than the choice of model**.

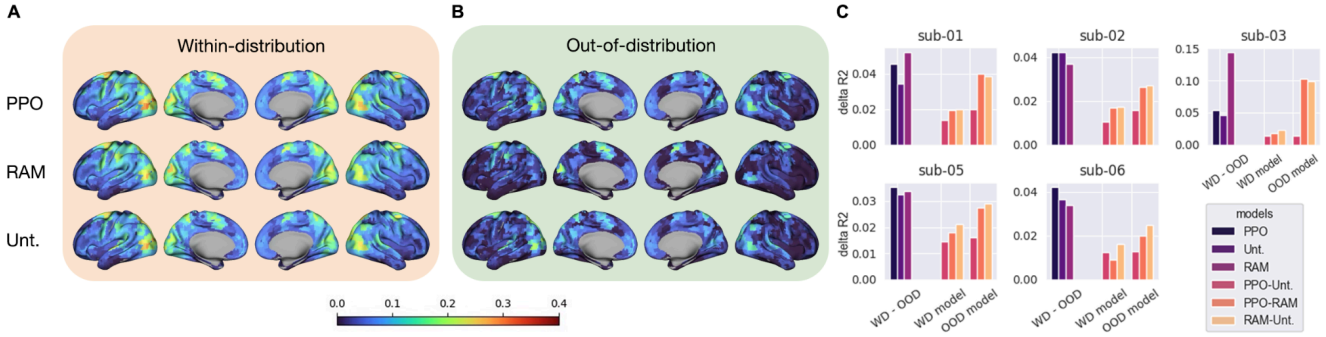

Figure 2: Brain encoding results. **A** and **B** show the results on within-distribution (WD) and out-of-distribution (OOD) testing, averaged across participants. **C** shows the changes in R2 when comparing WD to OOD testing (dark bars) relative to the changes observed when comparing models (light bars).

Discussion

All models achieved relatively high brain encoding scores under within-distribution (WD) conditions, with only modest differences observed between models. In contrast, out-of-distribution (OOD) scores were substantially lower, indicating that none of the models generalized well beyond their training distribution. These findings highlight the importance of evaluating brain encoding models in settings that go beyond the narrow confines of their training data. In this context, videogames provide an especially suitable paradigm: they offer rich and dynamic environments while still allowing for highly controlled variations across contexts and tasks. Our results demonstrate that, if brain scores are used to adjudicate between different models of brain function, the size and diversity of the stimulus set is critical to reach robust conclusions.

References

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*(1), 1–10. https://doi.org/10.1038/s42003-022-03036-1

Cross, L., Cockburn, J., Yue, Y., & O'Doherty, J. P. (2021). Using deep reinforcement learning to reveal how the brain encodes abstract state-space representations in high-dimensional environments. *Neuron*, *109*(4), 724-738.e7. https://doi.org/10.1016/j.neuron.2020.11.021

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S., Wright, J., Durnez, J., Poldrack, R., & Gorgolewski, K. J. (2018). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*. https://doi.org/10.1038/s41592-018-0235-4

Gifford, A. T., Bersch, D., St-Laurent, M., Pinsard, B., Boyle, J., Bellec, L., Oliva, A., Roig, G., & Cichy, R. M. (2025). *The Algonauts Project 2025 Challenge: How the Human Brain Makes Sense of Multimodal Movies* (No. arXiv:2501.00504). arXiv. https://doi.org/10.48550/arXiv.2501.00504

Gifford, A. T., Cichy, R. M., Naselaris, T., & Kay, K. (2025). *A 7T fMRI dataset of synthetic images for out-of-distribution modeling of vision* (No. arXiv:2503.06286). arXiv. https://doi.org/10.48550/arXiv.2503.06286

Huh, M., Cheung, B., Wang, T., & Isola, P. (2024). *The Platonic Representation Hypothesis* (No. arXiv:2405.07987). arXiv. https://doi.org/10.48550/arXiv.2405.07987

Nichol, A., Pfau, V., Hesse, C., Klimov, O., & Schulman, J. (2018). *Gotta Learn Fast: A New Benchmark for Generalization in RL* (No. arXiv:1804.03720). arXiv. https://doi.org/10.48550/arXiv.1804.03720

Nintendo. (1983). *Super Mario Bros.* [Computer software]. Nintendo.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). *Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?* (p. 407007). bioRxiv. https://doi.org/10.1101/407007

Shirakawa, K., Nagano, Y., Tanaka, M., Aoki, S. C., Majima, K., Muraki, Y., & Kamitani, Y. (2024). *Spurious reconstruction from brain activity* (No. arXiv:2405.10078). arXiv. https://doi.org/10.48550/arXiv.2405.10078

Urchs, S., Armoza, J., Moreau, C., Benhajali, Y., St-Aubin, J., Orban, P., & Bellec, P. (2019). MIST: A multi-resolution parcellation of functional brain networks. *MNI Open Research*, *1*(3), 3.

Data and code

The code used to generate this analysis is shared through the following repository: https://zenodo.org/records/15642065. Data request is available via the website https://www.cneuromod.ca/, or the Canadian Open Neuroscience Platform's portal (https://portal.conp.ca/) where 4 of the subjects have made their data publicly available.