MBM Essay 1: Human

Reverse-Engineering Brain Mechanisms through Interpretable Neural Networks

Bruno Sánchez Gómez

June 9, 2025

Part I
The Symbiotic Evolution of AI and Neuroscience: A Historical Perspective

1 Introduction

- Hook: The enduring quest to understand the human brain, the most complex computational machine known.
- Thesis Statement: The intertwined history of Artificial Intelligence and Neuroscience has led to a symbiotic relationship where advances in one field have consistently fueled progress in the other. This essay will trace this co-evolution, from early computational models inspired by neural processes to the contemporary use of interpretable neural networks to reverse-engineer the brain's intricate mechanisms, culminating in a review of the current state-of-the-art and future directions.
- Roadmap: Briefly outline the essay's structure, covering the historical review, the rise of deep learning, the advent of interpretable AI (XAI), and the application of these tools in modern neuroscience.

2 The Dawn of AI and Early Neural Models (1940s-1960s)

- **The McCulloch-Pitts Neuron (1943):** The first mathematical model of a biological neuron. Discuss its significance as a foundational concept for both AI and computational neuroscience.
- **Hebb's Postulate and Learning (1949):** "Neurons that fire together, wire together." Explain the Hebbian learning rule and its influence on early learning algorithms in neural networks.
- **The Perceptron (1958):** Frank Rosenblatt's invention and its initial promise for pattern recognition. Connect this to early models of sensory processing in the brain.
- **The "AI Winter" and its Thaw:** Briefly discuss the limitations identified by Minsky and Papert (1969) and the subsequent decline and eventual resurgence of neural network research.

3 Connectionism and the Rise of Parallel Distributed Processing (1980s)

- **The PDP Group and the "Connectionist" Bible:** Discuss the impact of Rumelhart, Hinton, and McClelland's work.
- **Backpropagation and Multi-Layer Networks:** Explain the significance of the backpropagation algorithm in training more complex networks, allowing for the modeling of more sophisticated cognitive functions.
- **Early Applications in Cognitive Science:** Provide examples of how these models were used to simulate and understand phenomena like language acquisition, memory, and perception.

4 The Deep Learning Revolution and its Impact on Neuroscience (2000s-Present)

- **The Unreasonable Effectiveness of Data and Computation:** The convergence of large datasets (e.g., ImageNet) and powerful GPUs.
- **Convolutional Neural Networks (CNNs) and the Visual Cortex:**
  - Draw strong parallels between the hierarchical structure of CNNs and the organization of the primate visual stream (V1, V2, V4, IT).
  - Discuss seminal work (e.g., Yamins & DiCarlo) showing that CNNs trained on object recognition tasks develop representations remarkably similar to those found in the visual cortex.
- **Recurrent Neural Networks (RNNs) and Sequential Processing:**
  - Explain the architecture of RNNs (including LSTMs and GRUs) and their suitability for modeling time-series data.
  - Connect this to the brain's processing of language, motor sequences, and decision-making over time.

Part II
The Black Box Dilemma and the Rise of Interpretable AI (XAI)

5 The "Black Box" Problem in Deep Learning

- **Defining the Challenge:** While deep neural networks achieve impressive performance, their internal workings are often opaque and difficult to understand.
- **Implications for Scientific Discovery:** In the context of neuroscience, a "black box" model that mimics brain function without revealing *how* it does so offers limited scientific insight. The goal is not just to replicate but to understand the underlying principles.

6 A Taxonomy of Interpretable AI Methods

- **Post-hoc vs. Intrinsically Interpretable Models:** Differentiate between methods that analyze a trained model and those that are designed to be transparent from the outset.
- **Feature Attribution Methods:**
  - **Saliency Maps:** Visualizing which input features (e.g., pixels in an image) are most important for a model's prediction.
  - **LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations):** Explain how these methods build simpler, interpretable local models to approximate the behavior of the complex "black box" model.
- **Model-based Interpretation:**
  - **Analyzing Network Activations and Representations:** Techniques for visualizing and understanding the features learned by different layers of a network.
  - **Causal Intervention and "Ablation" Studies:** Simulating lesion studies in neuroscience by deactivating specific neurons or connections in the network to observe the effect on its output.

7 Interpretable AI in Action: Bridging the Gap to Neuroscience

- **Reverse-Engineering Sensory Systems:**
  - **Vision:** Beyond object recognition, how XAI helps us understand how CNNs represent texture, shape, and motion, and how this compares to neural data from different visual areas.
  - **Audition:** Using deep learning models to understand the neural coding of sound, from the cochlea to the auditory cortex.
- **Decoding and Encoding Models:** Using interpretable models to predict neural activity from stimuli (encoding) and to decode mental states or perceived stimuli from neural recordings (decoding).
- **Understanding Higher Cognitive Functions:**

- **Language:** Analyzing the representations within large language models (LLMs) like BERT and GPT, and comparing them to the brain's language processing centers (e.g., Broca's and Wernicke's areas).
- **Decision-Making and Reinforcement Learning:** How interpretable reinforcement learning models can shed light on the neural circuits involved in reward, planning, and action selection.

Part III
The State of the Art and the Future of Brain-Inspired AI

8 Current Frontiers in Interpretable Neural Networks for Neuroscience

- **Generative Models and "In Silico" Experiments:** Using generative adversarial networks (GANs) and other generative models to create stimuli that maximally activate specific neurons or brain regions, allowing for more targeted experiments.
- **Beyond Supervised Learning:** The role of self-supervised and unsupervised learning in creating models that learn more brain-like representations without requiring massive labeled datasets.
- **The Rise of "Neuro-AI":** The growing field of research that explicitly aims to build AI systems based on principles from neuroscience, creating a virtuous cycle of discovery.
- **Thinking LLMs and Simulating Thought Processes:** Discussing the emerging use of large language models to model and understand human-like reasoning, planning, and problem-solving.

9 Challenges, Limitations, and Ethical Considerations

- **The "Simile" vs. "Model" Distinction:** Emphasize that even the most brain-like ANNs are still simplifications. Discuss the key biological details they often omit (e.g., dendritic computation, neuromodulation).
- **The Dangers of Over-interpretation:** The risk of drawing premature or overly simplistic conclusions about the brain based on analogies with AI models.
- **Data Privacy and Neuromarketing:** Briefly touch on the ethical implications of being able to decode brain states with increasing accuracy.

10 Conclusion: The Future of a Fruitful Partnership

- **Recap of the Main Arguments:** Summarize the historical co-evolution and the current state of synergy between AI and neuroscience.
- **Future Outlook:** Project how this interdisciplinary collaboration will continue to unravel the complexities of the brain and, in turn, inspire more general and capable artificial intelligence. The ultimate goal: a unified theory of intelligence, both biological and artificial.
- **Final Thought-Provoking Statement:** Reiterate the profound potential of this research to not only advance science but also to fundamentally alter our understanding of ourselves.

References

[1] Warren S McCulloch and Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4 (1943), pp. 115–133.

[2] Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, 1949.

[3] Frank Rosenblatt. "The perceptron: a probabilistic model for information storage and organization in the brain". In: *Psychological review* 65.6 (1958), p. 386.

[4] Marvin Minsky and Seymour Papert. *Perceptrons: An introduction to computational geometry*. MIT press, 1969.

[5] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536.

[6] James L McClelland, David E Rumelhart, and PDP Research Group. *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2*. MIT press, 1986.

[7] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[8] Daniel L Yamins and James J DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3 (2016), pp. 356–365.

[9] Nikolaus Kriegeskorte and Pamela K Douglas. "Cognitive computational neuroscience". In: *Nature Neuroscience* 21.9 (2018), pp. 1148–1160.

[10] Blake A Richards et al. "A deep learning framework for neuroscience". In: *Nature neuroscience* 22.11 (2019), pp. 1761–1770.

[11] Neil Savage. "How AI is helping to explain the brain". In: *Nature* 571.7766 (2019), S16–S18.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should I trust you?": Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[13] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems*. 2017, pp. 4765–4774.

[14] Chris Olah et al. "The building blocks of interpretability". In: *Distill* 3.3 (2018), e10.

[15] Finale Doshi-Velez and Been Kim. "Towards a rigorous science of interpretable machine learning". In: *arXiv preprint arXiv:1702.08608* (2017).

[16] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38.

[17] Patrick McClure and Nikolaus Kriegeskorte. "How to compare representations in brains and machines". In: *bioRxiv* (2024), pp. 2024–01.

[18] Ruth C Fong and Andrea Vedaldi. "Interpretable explanations of black boxes by meaningful perturbation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3429–3437.

[19] David Bau et al. "Network dissection: Quantifying interpretability of deep visual representations". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6541–6549.

[20] Martin Schrimpf et al. "Brain-Score: A framework for comparing computational models of the brain". In: *bioRxiv* (2020).

[21] Mariya Toneva and Leila Wehbe. "Interpreting and improving natural language processing models for neuro-imaging analyses". In: *Advances in Neural Information Processing Systems*. 2019, pp. 14382–14392.

[22] Charlotte Caucheteux and Jean-R'emi King. "Brains and algorithms partially converge in natural language processing". In: *Communications Biology* 5.1 (2022), pp. 1–12.

[23] Jason Wei et al. "Chain of thought prompting elicits reasoning in large language models". In: *arXiv preprint arXiv:2201.11903* (2022).

[24] Marcel Binz and Eric Schulz. "Using cognitive psychology to understand large language models". In: *Nature Reviews Psychology* 2.7 (2023), pp. 385–386.

[25] Kyle Mahowald et al. "Dissociating language and thought in large language models: a cognitive perspective". In: *arXiv preprint arXiv:2301.06627* (2023).