MBM Essay 1: Human

Reverse-Engineering Brain Mechanisms through Explainable AI

Bruno Sánchez Gómez

June 10, 2025

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

1 Introduction

The human brain is the most complex computational machine known to mankind. As such, philosophers and scientists have been fascinated by it for centuries, and have put great effort into trying to understand how its billions of neurons, working together, can create thoughts, perceptions, and consciousness.

This essay will look at the closely connected history of Artificial Intelligence (AI) and Neuroscience, two symbiotic fields that often take inspiration from each other. From the earliest attempts to create computational models based on neural processes to the cutting-edge use of explainable neural networks to better our understanding of the brain's complex mechanisms, progress in one domain has consistently spurred innovation in the other. We will follow this shared evolution, looking at how ideas from Neuroscience have influenced AI designs, and how AI now gives us new tools and ways to think about reverse-engineering the brain. The discussion will culminate in a review of the current state-of-the-art in this field and a look towards future directions where this collaboration promises to yield even deeper understanding of both biological and artificial intelligence.

The essay is structured as follows: Part I will begin with a historical review, charting the dawn of AI and early neural models, the rise of connectionism, and the transformative impact of the Deep Learning revolution on Neuroscience. Part II will delve into the "black box" problem inherent in many complex AI models and the subsequent birth of Explainable AI (XAI), detailing various methodologies and their crucial applications in closing the gap between AI performance and neuroscientific understanding. Finally, Part III will assess the current frontiers, discussing generative models, the shift beyond supervised learning, the rise of "Neuro-AI", and the use of large language models to simulate thought processes, while also considering the challenges, limitations, and ethical considerations inherent in this rapidly evolving field.

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

The foundational period of AI, spanning from the 1940s to the 1960s, was marked by pioneering efforts to conceptualize and model neural processes mathematically. A seminal contribution during this era was the McCulloch-Pitts neuron, introduced in 1943 [1]. This model presented a simplified, logical abstraction of a biological neuron, proposing that neurons could be understood as computational units performing logical operations. Although a simplification of true neuronal complexity, the McCulloch-Pitts neuron was a groundbreaking concept. It laid crucial groundwork for both the nascent field of AI, by suggesting that machines could, in principle, perform tasks analogous to human thought, and for computational Neuroscience, by providing a formal framework to begin modeling neural activity and networks.

Building on the idea of interconnected processing units, Donald Hebb, in his influential 1949 work, "The Organization of Behavior" [2], proposed a mechanism for learning in the brain. His famous postulate, often summarized as "neurons that fire together, wire together," introduced the concept of synaptic plasticity. Hebbian learning suggested that the strength of a connection between two neurons increases when they are activated simultaneously. This principle was profoundly influential, offering a plausible biological basis for how learning and memory could arise from neural activity and directly inspiring early learning algorithms in artificial neural networks. It provided a dynamic element to the static connections of earlier models, suggesting

how networks could adapt and learn from experience.

Furthering the development of learning machines, Frank Rosenblatt introduced the Perceptron in 1958 [3]. The Perceptron was a more concrete implementation of a learning neural network, capable of learning to classify patterns by adjusting its synaptic weights based on errors. This invention generated considerable excitement, as it demonstrated a machine that could learn from data and perform pattern recognition tasks, a key aspect of intelligence. The Perceptron's architecture and learning rule were seen as analogous to early models of sensory processing in the brain, particularly in how simple features might be detected and combined to recognize more complex stimuli. It represented a significant step towards building practical AI systems inspired by neural principles.

Despite these early successes and the initial optimism, the field soon encountered significant challenges. In 1969, Marvin Minsky and Seymour Papert published "Perceptrons" [4], a detailed mathematical analysis that highlighted severe limitations of single-layer Perceptrons, most notably their inability to solve problems that were not linearly separable, such as the XOR problem. This critique, coupled with exaggerated claims and unmet expectations, led to a significant reduction in funding and interest in neural network research, a period often referred to as the first "AI winter." However, the foundational ideas laid during these early decades were not entirely abandoned. They simmered beneath the surface, awaiting new algorithmic breakthroughs and computational power that would eventually lead to a resurgence and the development of more powerful, multi-layered neural networks in the decades to follow.

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

After the "AI winter," the 1980s witnessed a significant resurgence of interest in neural networks, largely fueled by the emergence of connectionism and the concept of Parallel Distributed Processing (PDP). Central to this revival was the work of the PDP Research Group, particularly David Rumelhart, Geoffrey Hinton, and James McClelland. Their influential two-volume book, "Parallel Distributed Processing: Explorations in the Microstructure of Cognition" [5], often referred to as the "Connectionist Bible," provided a comprehensive theoretical framework and compelling demonstrations of neural network capabilities. This work emphasized how complex cognitive phenomena could emerge from the parallel interactions of many simple processing units, akin to neurons. Connectionist models proposed that information was not stored in specific locations but distributed across connections, and that learning occurred through the modification of these connection strengths. This period also saw the popularization of ideas related to the capabilities of multi-layer networks.

A critical breakthrough that enabled the practical training of these more complex, multi-layer networks was the popularization and refinement of the backpropagation algorithm, notably described by Rumelhart, Hinton, and Williams in 1986 [6]. While the core ideas of backpropagation had been explored earlier by others, their work made it accessible and demonstrated its power. Backpropagation provided an efficient method for calculating the gradient of the error function with respect to the network's weights, allowing for the systematic adjustment of these weights to minimize error. This meant that networks with hidden layers, which were previously difficult to train, could now learn complex input-output mappings. The ability to train multi-layer networks was a game-changer, as these architectures could overcome the limitations of single-layer perceptrons identified by Minsky and Papert, enabling the modeling of non-linear relationships and, consequently, more sophisticated cognitive functions.

The advancements in network architectures and training algorithms led to a wave of early applications in cognitive science, where PDP models were used to simulate and offer insights into a variety of human cognitive processes. These early applications, while often simplified, were crucial in demonstrating the potential of neural networks as tools for understanding the mind

and brain, bridging the gap between abstract computational principles and observable cognitive phenomena [5].

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

The new millennium marked the beginning of a new era for AI, largely characterized by the rise of Deep Learning. This revolution, which has only gained more momentum since the early 2000s, has had a profound and reciprocal impact on Neuroscience. The convergence of several key factors led to what some have termed the "unreasonable effectiveness" of Deep Learning models [7].

- Firstly, the availability of massive datasets, such as ImageNet, provided the rich training material necessary for complex models to learn intricate patterns.
- Secondly, the parallel development and widespread adoption of powerful Graphics Processing Units (GPUs) offered the computational horsepower required to train these data-hungry and computationally intensive deep neural networks in a feasible timeframe.
- This combination of big data and powerful hardware unlocked the potential of Deep Learning architectures, allowing them to achieve state-of-the-art performance on a wide range of tasks, from image recognition to natural language processing, and in doing so, provided new, powerful tools for neuroscientific inquiry [8].

Among the most influential Deep Learning architectures for Neuroscience have been Convolutional Neural Networks (CNNs). Initially inspired by the hierarchical processing observed in the mammalian visual cortex, CNNs have demonstrated remarkable success in computer vision tasks. These networks typically consist of multiple layers, including: convolutional layers, that apply filters to input images to detect features; pooling layers, that reduce dimensionality; and fully connected layers, that perform classification. The hierarchical nature of CNNs, where early layers learn simple features (like edges and textures), and deeper layers learn more complex and abstract representations (like object parts or even whole objects), bears a striking resemblance to the processing stages in the primate visual pathway [9]. This architectural similarity has made CNNs invaluable tools for modeling the visual cortex. Neuroscientists have used them not only to predict neural responses to visual stimuli with unprecedented accuracy but also to generate hypotheses about how visual information is represented and transformed in the brain [10, 11]. By comparing the internal representations of CNNs trained on visual tasks with neural recordings from different stages of the visual system, researchers have found compelling correspondences, suggesting that these artificial models might capture fundamental principles of biological vision.

Similarly, Recurrent Neural Networks (RNNs) have provided powerful models for understanding how the brain processes sequential data, such as language or temporal patterns in sensory input. Unlike feedforward networks like CNNs, RNNs possess connections that loop back on themselves, allowing them to maintain an internal state or "memory" of past inputs. This characteristic makes them well-suited for tasks where context and temporal dependencies are crucial. The ability of RNNs, particularly variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) which can learn long-range dependencies, to capture complex temporal dynamics has made them instrumental in exploring the neural mechanisms underlying time-sensitive cognitive processes [8].

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

5 The "Black Box" Problem in Deep Learning

The remarkable success of Deep Learning models in emulating complex cognitive functions [7], as explored in Section 4, brings with it a significant challenge: their inherent opacity. Deep neural networks, with their vast number of parameters and intricate, multi-layered architectures, often operate as "black boxes" [12]. While their input-output behavior can be meticulously evaluated, the internal computations and learned representations that lead to a specific decision or prediction are frequently obscure and difficult for humans to comprehend directly. This lack of transparency means that even when a model achieves high performance, understanding *why* it performs well or *how* it arrives at its conclusions can be an arduous task, undermining trust and hindering deeper scientific inquiry [13, 14].

This "black box" characteristic poses a particular impediment when these models are applied to Neuroscience, where the primary objective extends beyond mere replication of brain-like outputs to a fundamental understanding of the underlying neural mechanisms [8, 11]. If a Deep Learning model accurately predicts neural responses in the visual cortex [9] or simulates language processing [15] but its internal strategies remain inscrutable, it offers limited insight into how the biological brain actually performs these tasks. The aspiration of cognitive computational Neuroscience is to develop models that are not only predictive but also explanatory, providing testable hypotheses about brain function [10]. A model that functions as an uninterpretable oracle, however accurate, falls short of this goal. It may demonstrate that a certain mapping from stimulus to neural activity or behavior is learnable, but it fails to reveal the computational principles or representational transformations that the brain itself might employ. Consequently, the scientific value for reverse-engineering brain mechanisms is obscured, as it becomes challenging to validate whether the model's learned solutions are neurobiologically plausible or to derive new insights into the brain's algorithms. The critical need, therefore, is to develop approaches that can illuminate these internal workings, transforming powerful predictive tools into genuinely explanatory frameworks for understanding the brain.

6 A Taxonomy of Explainable AI Methods

To address the opacity of Deep Learning models, the field of Explainable AI (XAI) has emerged, offering a diverse variety of methods to take a look inside the "black box" [12, 14]. A primary distinction in XAI methodologies lies between *post-hoc interpretability techniques*, which are applied to already trained models, and *intrinsically explainable models*, which are designed for transparency from the ground up. While some models like linear regression or decision trees are inherently interpretable due to their simple structure, the complex, highly non-linear nature of deep neural networks means they seldom fall into this category. Consequently, much of XAI research focuses on post-hoc methods to analyze these powerful but opaque systems [12].

One prominent category of post-hoc XAI techniques comprises **Feature Attribution Methods**. These methods aim to identify which parts of the input data a model deems most important when making a specific prediction.

- **Saliency Maps** are a common feature attribution technique, particularly in computer vision. They generate heatmaps that highlight the pixels in an input image that most significantly influence the model's output for a given class [16]. Typically, these are computed by examining the gradient of the output prediction score with respect to the input

pixel values. For neuroscientists, saliency maps can offer a visual hypothesis about which features a model, trained to mimic a sensory processing task, might be using, analogous to how one might study receptive fields in biological neurons.

- **LIME (Local Interpretable Model-agnostic Explanations)** and **SHAP (SHapley Additive exPlanations)** offer more sophisticated approaches to feature attribution. LIME operates by training a simpler, interpretable model (e.g., a linear model or a decision tree) to approximate the behavior of the complex black-box model in the local vicinity of a particular instance being explained [13]. This provides a local, understandable rationale for a specific prediction without requiring insight into the global structure of the original model. SHAP, on the other hand, draws from cooperative game theory, specifically Shapley values, to assign a unique contribution value to each feature for a given prediction, providing a unified and theoretically grounded framework for feature importance [17]. These methods are valuable because they can be applied to virtually any model (model-agnostic) and help study the reasoning behind individual predictions, which is crucial when trying to understand if an AI model's strategy bears any resemblance to neural computation.

Beyond attributing importance to input features, another class of XAI techniques focuses on **Model-based Interpretation**, aiming to understand the internal mechanisms and learned representations within the network itself.

- **Analyzing Network Activations and Representations** involves directly inspecting the internal state of the network. This can include visualizing the activation patterns of individual neurons or entire layers in response to different inputs, or using techniques like feature visualization, which attempts to synthesize an input that maximally activates a specific neuron or feature detector within the network [18]. Such methods allow researchers to probe what kinds of information are encoded at different stages of processing. For instance, the hierarchical features that CNNs learn when trained on image recognition, as we mentioned in Section 4 [9]. Other tools, like Network Dissection [19], aim to quantify the interpretability of individual units by identifying the semantic concepts (e.g., 'door', 'sky') that they detect.

- **Ablation Studies** in artificial neural networks are directly inspired by lesion studies in Neuroscience. These methods involve systematically perturbing the model (for example, by desactivating specific artificial neurons, connections, or entire layers) and then observing the impact on the model's behavior or internal representations [10]. If silencing a particular component significantly degrades performance on a specific task, it suggests that this component plays a crutial role in that computation. Such "in silico lesioning" can help to map functional roles to different parts of the network, providing testable hypotheses about the functional specialization of the different regions of the network.

These various XAI approaches are not mutually exclusive and are often used in combination to build a more comprehensive understanding of how complex AI models operate, thereby offering a richer set of tools for comparison with, and generation of hypotheses about, the brain.

7 Explainable AI in Action: Bridging the Gap to Neuroscience

The theoretical tools of XAI, as outlined in Section 6, are not merely abstract concepts; they are actively being applied to bridge the gap between the high performance of AI models and a deeper neuroscientific understanding. By moving beyond simple performance metrics and towards model interpretability, researchers can start to reverse-engineer the complex computational processes of the brain.

One of the most fruitful applications of XAI in Neuroscience has been in the domain of **reverse-engineering sensory systems**. In the field of vision, for instance, while the initial

success of CNNs was in object recognition, XAI techniques have allowed for a more nuanced analysis of their internal workings. Feature attribution methods and the analysis of network activations have revealed how these models learn to represent not just whole objects, but also fundamental visual properties like texture, shape, and motion. By comparing these learned representations with neural data from different stages of the visual cortex, researchers can generate and test specific hypotheses about how the brain deconstructs and processes visual scenes [9, 10]. Similarly, in audition, Deep Learning models are being used to understand the neural coding of sound. Explainable models help to uncover how features like pitch, timbre, and rhythm are extracted and represented in the hierarchical layers of these networks, offering insights into the transformations that might be occurring in the biological auditory pathway [8].

XAI methods are also central to the development of more sophisticated **encoding and decoding models** for brain signals. Encoding models aim to predict neural activity from given stimuli, while decoding models attempt the reverse: to infer mental states or perceived stimuli from neural recordings. Explainable models are crucial here, as they allow researchers to understand *which* features of a stimulus drive neural responses (in encoding) or *which* patterns of neural activity are most informative for a particular mental state (in decoding). For example, an explainable encoding model for the visual cortex might reveal not just that a certain neuron fires in response to a face, but that its firing is specifically driven by the presence of an eye or the curvature of a smile. This level of detail is essential for building a mechanistic understanding of neural representation and for developing more precise brain-computer interfaces [20].

Furthermore, XAI is beginning to shed light on **higher cognitive functions**. In the realm of language, researchers are now probing the internal representations of large language models (LLMs) like BERT and GPT to understand how they process and represent linguistic information. By comparing the activation patterns within these models to brain imaging data from humans engaged in language tasks, studies have found intriguing parallels between the models' representations and the activity in classical language centers of the brain, such as Broca's and Wernicke's areas [15, 21]. This comparative approach helps to formulate hypotheses about how the brain might implement computations for syntax, semantics, and context. In the domain of decision-making, explainable reinforcement learning (RL) models are providing insights into the neural circuits underlying reward, planning, and action selection. By analyzing the value functions and policies learned by RL agents, researchers can draw parallels to the dopaminergic systems and prefrontal cortex activity involved in goal-directed behavior, offering a computational framework for understanding both adaptive and maladaptive decision-making processes [8]. Through these applications, XAI is proving to be an indispensable tool, transforming AI models from black-box mimics into interpretable frameworks for generating and testing hypotheses about the brain.

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

8 Current Frontiers in Explainable Neural Networks for Neuroscience

The XAI tools and approaches detailed in the previous part are essential for interpreting trained AI models. Current frontiers, however, are not only about post-hoc explanation but also involve developing AI in ways that are intrinsically more aligned with neuroscientific inquiry or that enable novel experimental paradigms. One such frontier is the sophisticated use of generative models for "in silico" experiments. Generative models, such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), learn the underlying distribution of data and can synthesize new data samples. In the context of Neuroscience, these models are increasingly employed to create optimized or novel stimuli designed to maximally activate specific artificial neurons within a neural network model, or even to predict stimuli that would maximally drive responses in biological neurons or entire brain regions [18]. This technique, often referred to as activation maximization or feature visualization, allows researchers to go beyond merely observing what features a model has learned and actively probe what specific patterns or combinations of features are most salient for particular units or representations. For instance, if a model of the visual cortex learns a specific representation, generative methods can synthesize the "ideal" visual input for that representation, providing a concrete, visual hypothesis about its tuning properties. These synthetic stimuli can then, in turn, be used in actual neurophysiological experiments, creating a powerful feedback loop between computational modeling and empirical investigation [8, 10]. This allows for more targeted and efficient experimentation than relying solely on pre-defined stimulus sets, potentially uncovering unexpected feature selectivity.

While generative models provide powerful means to probe the learned representations of neural networks, another critical frontier concerns how such rich, potentially brain-like representations are acquired, especially considering the learning constraints of biological systems. This leads to investigations beyond predominantly supervised learning paradigms. The brain, particularly during early development, excels at learning from largely unlabeled or sparsely labeled data. Consequently, there is growing interest in self-supervised learning (SSL) and unsupervised learning (UL) for creating AI models that might develop more brain-like representations [8]. SSL methods, for example, create supervisory signals from the data itself (e.g., by predicting a missing part of an input, or learning to be invariant to certain transformations), while UL aims to discover inherent structure, like clusters or principal components, in the data without any explicit labels. Models trained with these paradigms are hypothesized to capture the statistical regularities of their input environment in a manner more akin to how biological sensory systems operate. The resulting representations can be more robust, generalizable, and potentially more aligned with the rich, multifaceted representations found in the brain, which are not solely optimized for a single, narrowly defined task. Exploring these learning frameworks is crucial for developing AI that not only performs well but also learns in a way that could offer deeper insights into the principles of neural learning and development.

The development of models using self-supervised or unsupervised learning, which aim for more biologically plausible representational learning, is indicative of a broader trend: the rise of "Neuro-AI." This research field is explicitly dedicated to fostering a deeper, synergistic relationship between Neuroscience and Artificial Intelligence, moving beyond unidirectional inspiration towards a virtuous cycle of discovery [8, 10]. In this paradigm, neuroscientific findings about brain architecture (e.g., distinct processing pathways, recurrent connectivity), learning rules

(e.g., Hebbian plasticity, synaptic consolidation), and computational principles (e.g., predictive coding, sparse representations, attention mechanisms) directly inform the design of new AI architectures and algorithms. These more biologically-inspired or constrained AI systems are then leveraged not only for technological advancement but also as more plausible and testable computational models of brain function. By meticulously comparing the behavior and internal dynamics of these Neuro-AI models with neural and behavioral data, researchers can refine their understanding of the brain. This, in turn, generates new, empirically testable hypotheses that can feed back into the development of even more sophisticated AI, ultimately aiming for models that are both performant and mechanistically interpretable in a way that resonates with biological reality.

The ambitions of Neuro-AI, seeking to bridge the gap between artificial systems and biological intelligence, find a particularly compelling and rapidly evolving testbed in the study of Large Language Models (LLMs) and their potential to simulate complex thought processes. These models have demonstrated remarkable abilities in processing and generating human language, and more recently, in tasks that seem to require reasoning, planning, and problem-solving [22]. The emergence of such capabilities has spurred intense investigation into whether and how these models might be "simulating" or even instantiating aspects of human thought. XAI techniques are vital in this endeavor, as researchers attempt to peer inside these vast networks to understand the basis of their performance. For example, studies analyzing the effect of "chain-of-thought" prompting, where models are encouraged to output intermediate reasoning steps, suggest that LLMs can indeed follow and articulate a form of step-by-step reasoning [22]. Furthermore, by comparing the internal representational spaces of LLMs with brain activity patterns recorded while humans perform language tasks, researchers have found intriguing alignments, suggesting some convergence in how linguistic information is structured [15, 21]. However, this is an area of active debate and research, particularly concerning the extent to which these abilities reflect genuine understanding versus sophisticated pattern matching, and how language capabilities relate to broader cognitive thought [23, 24]. Investigating LLMs through a cognitive lens, including their operational mechanisms, limitations, and biases, offers a novel platform for generating hypotheses about the mechanisms of human reasoning, language comprehension, and their intricate interplay.

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

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

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

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

[9] 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.

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

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

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

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

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

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