MBM Essay 2: LLM

From Neural Inspiration to Neural Explanation: The Symbiotic Journey of AI and Neuroscience

Bruno Sánchez Gómez

Content written by Gemini 2.5 Pro

June 11, 2025

From Neural Inspiration to Neural Explanation: The Symbiotic Journey of AI and Neuroscience

A Language Model

June 11, 2025

Abstract

The relationship between Artificial Intelligence (AI) and Neuroscience is one of the most profound and enduring intellectual partnerships in modern science. Born from a shared curiosity about the mechanisms of intelligence, these fields have embarked on a symbiotic journey, oscillating between periods of close collaboration and independent exploration. Initially, neuroscience provided the primary inspiration for AI, with early computational models directly attempting to emulate neural structures. As AI evolved, particularly with the advent of deep learning, it has not only achieved remarkable engineering feats but has also offered increasingly sophisticated tools and conceptual frameworks for understanding the brain itself. This essay charts the historical co-evolution of AI and neuroscience, from early biologically-inspired models to the current era where advanced AI systems serve as testbeds for neuroscientific theories. We then delve into the burgeoning field of Explainable AI (XAI), examining its critical role in bridging the gap between complex AI models and human understanding, and how these methodologies are being adapted and applied to unravel the intricate workings of biological neural circuits. The current state-of-the-art in XAI for neuroscience is explored, highlighting its potential to transform our comprehension of neural computation, alongside a critical analysis of its inherent challenges, limitations, and the pressing ethical considerations that accompany this powerful new frontier. Ultimately, we argue that the journey from neural inspiration to neural explanation represents a maturing of the AI-neuroscience relationship, poised to unlock deeper insights into both artificial and biological intelligence.

Contents

1 Introduction

The quest to understand intelligence, whether embodied in biological organisms or engineered into artificial systems, stands as one of humanity's grandest scientific and philosophical challenges. At the heart of this pursuit lie two distinct yet deeply intertwined disciplines: Neuroscience, the empirical study of nervous systems and the brain, and Artificial Intelligence (AI), the endeavor to create machines that can perform tasks typically requiring human intelligence. Their relationship, from its inception, has been characterized by a dynamic interplay of inspiration, divergence, and renewed convergence, a journey that has profoundly shaped the trajectory of both fields.

Initially, the nascent field of AI drew heavily upon the then-emerging understanding of the brain's architecture. The neuron, as the fundamental computational unit of the brain, became the muse for early AI pioneers, leading to the development of artificial neural networks (ANNs). These models, though vastly simplified, sought to capture the essence of neural processing. For decades, this "neural inspiration" fueled AI research, even as the complexities of the brain often outpaced the capabilities of computational models and theoretical frameworks.

However, the tide has begun to turn. The recent and spectacular successes of deep learning, a subfield of AI employing multi-layered ANNs, have not only revolutionized technology but have also provided neuroscientists with powerful new tools and, perhaps more importantly, testable hypotheses about neural function. Complex AI models, trained on vast datasets to perform tasks like image recognition or natural language processing, exhibit internal representations and computational strategies that, in some cases, bear striking resemblance to those observed in biological brains. This has opened a new chapter in the AI-neuroscience saga: the shift from AI merely being *inspired by* the brain to AI *helping to explain* the brain.

This transition is critically dependent on our ability to understand the inner workings of these sophisticated AI models. Many state-of-the-art AI systems, particularly deep neural networks, operate as "black boxes"—their decision-making processes are opaque even to their creators. This opacity limits their scientific utility as models of the brain and raises concerns about their reliability and trustworthiness in critical applications. Herein lies the importance of Explainable AI (XAI), a rapidly developing area of research focused on creating techniques to render AI models more transparent and interpretable.

This essay embarks on an exploration of this "symbiotic journey," tracing the path from neural inspiration to the promise of neural explanation.

- We will begin with a historical review of the joint evolution of AI and neuroscience, highlighting key milestones and paradigm shifts that have defined their relationship.
- We will then conduct an in-depth analysis of Explainable AI, discussing its motivations, core concepts, and methodologies, and specifically how it forges a crucial link to neuroscience.
- Subsequently, we will examine the current state-of-the-art applications of XAI in neuroscience, showcasing how these tools are being leveraged to probe neural data and gain insights into brain function.
- Critically, we will also address the significant challenges, inherent limitations, and pressing ethical considerations associated with using AI, and XAI in particular, to explain the brain.

By navigating this complex landscape, we aim to illuminate how the confluence of AI and neuroscience, augmented by the drive for explainability, is not just pushing the boundaries of what machines can do, but is fundamentally reshaping our approach to understanding the most complex known object in the universe: the human brain. The journey is far from over, but the prospect of AI facilitating a deeper "neural explanation" signifies a pivotal maturation in this enduring intellectual partnership.

2 A Historical Tapestry: The Intertwined Evolution of AI and Neuroscience

The relationship between AI and neuroscience is not a recent phenomenon but a long-standing dialogue, marked by periods of intense cross-pollination, mutual inspiration, and occasional divergence. Understanding this historical context is crucial for appreciating the current state and future potential of their synergy.

2.1 Early Seeds of Inspiration (Pre-1950s)

The conceptual groundwork for AI and computational neuroscience was laid even before the advent of digital computers.

- **The Neuron Doctrine and Early Computational Ideas:** Santiago Ramón y Cajal's work in the late 19th and early 20th centuries established the neuron as the fundamental structural and functional unit of the nervous system [Ramón y Cajal, 1894]. This "neuron doctrine" provided a discrete, cellular basis for thinking about brain computation.
- **McCulloch and Pitts' Logical Calculus (1943):** A landmark paper by Warren McCulloch, a neurophysiologist, and Walter Pitts, a logician, proposed the first mathematical model of a neuron [McCulloch & Pitts, 1943]. Their model depicted neurons as simple binary devices performing logical operations (AND, OR, NOT). They demonstrated that networks of these "formal neurons" could, in principle, compute any function computable by a Turing machine. This was a pivotal moment, explicitly linking neural activity to computation and laying a theoretical foundation for artificial neural networks.
- **Hebb's Postulate and Learning (1949):** Donald Hebb, a psychologist, proposed a physiological mechanism for learning and memory in his book "The Organization of Behavior" [Hebb, 1949]. His famous postulate, "neurons that fire together, wire together," suggested that synaptic strength between neurons increases when they are persistently co-active. Hebbian learning provided a plausible biological rule for how neural networks could adapt and learn from experience, influencing AI learning algorithms for decades.
- **Wiener and Cybernetics (1948):** Norbert Wiener's work on cybernetics, "Cybernetics: Or Control and Communication in the Animal and the Machine," explored concepts of feedback, control, and information processing in both biological and artificial systems [Wiener, 1948]. This provided a broader intellectual framework that encompassed both neuroscience and the nascent ideas of intelligent machines.

These early contributions, rooted in observations about biological nervous systems, provided the intellectual kindling for the formal birth of AI.

2.2 The Dawn of AI and Early Neural Networks (1950s-1970s)

The 1950s witnessed the formal emergence of AI as a distinct field, with neural inspiration playing a central role.

- **The Dartmouth Workshop (1956):** Often cited as the birthplace of AI as a field, this workshop, organized by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, brought together researchers to explore the conjecture that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." Early neural network concepts were part of this discussion.
- **The Perceptron (1958):** Frank Rosenblatt, a psychologist, developed the Perceptron, an early single-layer neural network [Rosenblatt, 1958]. It could learn to classify patterns by adjusting its

synaptic weights based on an error-correction rule. The Perceptron generated immense excitement, demonstrating that simple neural-like structures could perform learning tasks.

- **ADALINE and MADALINE (1960s):** Bernard Widrow and Tedd Hoff developed ADALINE (Adaptive Linear Neuron) and MADALINE (Multiple ADALINE), which used a different learning rule (the LMS algorithm, or Widrow-Hoff rule) and found practical applications in areas like adaptive filtering [Widrow & Hoff, 1960].
- **Early Optimism and Biological Fidelity:** During this period, there was a strong belief that understanding the brain was key to building AI, and conversely, that building AI would illuminate brain function. Models were often directly compared to biological structures.
- **The "AI Winter" and Minsky & Papert's Critique (1969):** The initial enthusiasm waned. In their influential book "Perceptrons," Marvin Minsky and Seymour Papert rigorously analyzed the limitations of single-layer perceptrons, proving they could not solve certain classes of problems (e.g., the XOR problem) [Minsky & Papert, 1969]. This critique, coupled with unmet expectations and reduced funding, led to the first "AI winter," particularly for neural network research. While their mathematical analysis was sound for single-layer networks, it inadvertently dampened enthusiasm for multi-layer networks, which could overcome these limitations.

During this "winter," symbolic AI, focusing on logic, rule-based systems, and knowledge representation, became the dominant paradigm in AI, moving away from direct neural inspiration.

2.3 The Re-emergence: Connectionism and Parallel Distributed Processing (1980s)

The 1980s saw a resurgence of interest in neural networks, under the banner of "connectionism" or "Parallel Distributed Processing" (PDP).

- **Hopfield Networks (1982):** John Hopfield introduced a type of recurrent neural network that could serve as associative memory and solve optimization problems [Hopfield, 1982]. These networks had well-defined dynamics and demonstrated how collective behavior in a network could lead to emergent computational properties, reigniting interest in neural computation.
- **The Backpropagation Algorithm (Mid-1980s):** The popularization of the backpropagation algorithm by David Rumelhart, Geoffrey Hinton, and Ronald Williams was a watershed moment [Rumelhart, Hinton, & Williams, 1986]. While the core ideas existed earlier (e.g., Werbos, 1974), their work demonstrated its effectiveness for training multi-layer perceptrons (MLPs). Backpropagation provided an efficient way to assign credit (or blame) to hidden units in a network, enabling MLPs to learn complex, non-linear mappings and overcome the limitations highlighted by Minsky and Papert.
- **The PDP Books (1986):** The two-volume book "Parallel Distributed Processing: Explorations in the Microstructure of Cognition" by Rumelhart, McClelland, and the PDP Research Group became the bible of connectionism [McClelland, Rumelhart, & PDP Research Group, 1986]. It showcased numerous examples of how connectionist models could account for a wide range of psychological phenomena, from perception and memory to language.
- **Neuroscience Links:** Connectionist models often aimed for cognitive plausibility and were compared to brain processes. For example, work on self-organizing maps by Teuvo Kohonen provided models for topographic map formation in the cortex [Kohonen, 1982].

This period marked a strong re-engagement between AI and cognitive science, with neural networks providing a unifying framework.

2.4 Divergence and Specialization (1990s - Early 2000s)

While connectionism continued to develop, mainstream AI in the 1990s and early 2000s saw a shift towards more statistical and machine learning approaches that were less directly inspired by neuroscience.

- **Rise of Statistical Machine Learning:** Techniques like Support Vector Machines (SVMs) [Cortes & Vapnik, 1995], decision trees, and probabilistic graphical models gained prominence. These methods often had strong mathematical foundations and achieved excellent performance on many tasks, without necessarily claiming biological plausibility.
- **Computational Neuroscience Matures:** Simultaneously, computational neuroscience solidified as a distinct field. It focused on building more biologically realistic models of neurons and neural circuits, often incorporating detailed biophysical properties and anatomical connectivity. Researchers like David Marr laid out influential frameworks for understanding information processing in the brain at different levels of analysis (computational theory, representation and algorithm, hardware implementation) [Marr, 1982].
- **Continued, but Niche, Interaction:** While the mainstream of AI and neuroscience pursued somewhat separate paths, interaction continued. For example, reinforcement learning, with its roots in animal learning theory (e.g., temporal difference learning [Sutton & Barto, 1998]), found strong parallels with dopamine signaling in the basal ganglia [Schultz, Dayan, & Montague, 1997]. Models of visual cortex, like HMAX [Riesenhuber & Poggio, 1999], built upon hierarchical processing ideas inspired by Hubel and Wiesel's discoveries [Hubel & Wiesel, 1962].

This era was characterized by specialization, with AI focusing on engineering performance and computational neuroscience on biological realism, leading to a partial divergence.

2.5 The Deep Learning Renaissance and Renewed Symbiosis (Late 2000s - Present)

The late 2000s and 2010s witnessed a dramatic resurgence of neural networks, now rebranded as "deep learning," leading to a powerful new phase of AI-neuroscience interaction.

- **Key Drivers for Deep Learning:** Several factors converged:
  - **Big Data:** The availability of massive labeled datasets (e.g., ImageNet [Deng et al., 2009]).
  - **GPU Computing:** The use of Graphics Processing Units (GPUs) for parallel computation dramatically accelerated the training of large networks.
  - **Algorithmic Advances:** New activation functions (e.g., ReLU), regularization techniques (e.g., dropout), improved optimization algorithms, and novel architectures (e.g., Convolutional Neural Networks (CNNs) [LeCun et al., 1998], Recurrent Neural Networks (RNNs) like LSTMs [Hochreiter & Schmidhuber, 1997], and Transformers [Vaswani et al., 2017]).
- **Breakthrough Performance:** Deep learning models achieved state-of-the-art results on a wide array of challenging tasks, including image recognition (e.g., AlexNet [Krizhevsky, Sutskever, & Hinton, 2012]), speech recognition, natural language processing, and game playing (e.g., AlphaGo [Silver et al., 2016]).
- **ANNs as Models *of* the Brain:** A crucial shift occurred. Previously, ANNs were mostly *inspired by* the brain. Now, high-performing deep neural networks, particularly CNNs trained for visual object recognition, began to be seriously considered as quantitative models *of* information processing in the visual cortex [Yamins et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014]. Researchers found that the internal representations learned by these networks showed striking similarities to neural activity patterns recorded from different stages of the primate visual pathway.

- **Neuroscience Informing AI (and vice-versa):** This renewed connection became a two-way street.
  - **AI for Neuroscience Tools:** AI techniques are increasingly used to analyze complex neuroscience data (e.g., automated image segmentation, spike sorting, decoding neural activity).
  - **Neuroscience for AI Inspiration (Again):** Concepts from neuroscience, such as attention mechanisms [Bahdanau, Cho, & Bengio, 2014] (inspired by visual attention in humans) and memory systems, are being incorporated into AI architectures. The pursuit of more general and robust AI continues to draw inspiration from the brain's efficiency and adaptability.
- **The Rise of "NeuroAI":** This renewed synergy has led to the emergence of a vibrant interdisciplinary field, sometimes referred to as "NeuroAI," focused on leveraging insights and methodologies from both domains to advance the understanding of intelligence in both biological and artificial forms.

The historical journey has thus come full circle, but to a much higher plane. The initial, somewhat naive, neural inspiration has evolved into a sophisticated interplay where complex AI models serve not just as engineering tools but as functional hypotheses about brain computation, testable through rigorous neuroscientific experimentation and analysis, often aided by XAI.

3 Explainable AI (XAI): Bridging the Black Box to Understanding

As Artificial Intelligence systems, particularly deep neural networks (DNNs), have grown in complexity and predictive power, they have also become increasingly opaque. These "black box" models can make highly accurate predictions or decisions, but the internal reasoning or mechanisms leading to those outputs are often hidden from human users and even their developers. This lack of transparency poses significant challenges for trust, debugging, scientific discovery, and ethical deployment. Explainable AI (XAI) has emerged as a critical subfield of AI dedicated to developing methods that make AI systems more interpretable and understandable to humans.

3.1 The Need for Explainability

The demand for XAI stems from several pressing needs across various domains:

- **Trust and Reliability:** For AI systems to be adopted in high-stakes applications (e.g., medical diagnosis, autonomous driving, financial decision-making), users must trust their outputs. Explanations can help build this trust by revealing the basis for AI decisions, allowing users to assess their validity and reliability.
- **Debugging and Model Improvement:** When an AI model makes an error, understanding why it failed is crucial for debugging and improving its performance. XAI techniques can help identify flaws in the model's logic, biases in the training data, or unexpected failure modes.
- **Fairness, Accountability, and Transparency (FAT):** AI models can inadvertently learn and perpetuate societal biases present in their training data. XAI can help uncover these biases, promoting fairness. It also supports accountability by making it possible to trace and understand decision-making processes, which is essential for regulatory compliance (e.g., GDPR's "right to explanation").
- **Scientific Discovery:** In scientific domains like neuroscience, biology, or physics, AI models are increasingly used to analyze complex data and discover new patterns. XAI can help translate the opaque findings of these models into human-understandable insights, potentially leading to new scientific hypotheses and knowledge. This is particularly relevant for the theme of this essay.

- **Human-AI Collaboration:** For humans to effectively collaborate with AI systems, they need to understand the AI's strengths, weaknesses, and reasoning. Explanations can facilitate more effective human-AI teaming.

3.2 Defining Explainable AI: What Constitutes an "Explanation"?

There is no single, universally accepted definition of "explainability" or "interpretability." These terms are often used interchangeably, though some distinctions can be made (e.g., interpretability as the passive characteristic of a model being understandable, while explainability is the active process of providing an explanation). An "explanation" itself can take many forms, depending on the target audience, the model, and the context. Key characteristics of good explanations often include:

- **Fidelity:** The explanation should accurately reflect the model's reasoning process.
- **Intelligibility:** The explanation must be understandable to the intended human user. This means using concepts and language familiar to the user.
- **Actionability:** The explanation should provide insights that allow the user to take meaningful action (e.g., correct the model, make a decision).
- **Generality vs. Specificity:** Some explanations aim to provide a global understanding of the model's overall behavior, while others focus on explaining specific individual predictions.
- **Completeness:** While full transparency might be overwhelming, the explanation should capture the most salient factors influencing the model's output.

The "right" level and type of explanation are highly context-dependent. A machine learning engineer debugging a model might need a different explanation than a doctor using an AI for diagnostic support, or a neuroscientist trying to understand brain function via an AI model.

3.3 Key XAI Techniques

XAI methods can be broadly categorized in several ways, such as by scope (global vs. local), model-specificity (agnostic vs. specific), or the type of explanation produced. Here are some prominent categories and examples:

3.3.1 Feature Attribution Methods

These methods aim to identify which input features were most important for a particular prediction.

- **Saliency Maps / Sensitivity Analysis:** For image models, these visualize which pixels in an input image most influence the output classification [Simonyan, Vedaldi, & Zisserman, 2013]. This is often done by computing the gradient of the output score with respect to the input pixels. Variants include SmoothGrad, Grad-CAM, and Guided Backpropagation.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME explains individual predictions of any black-box classifier by learning a simpler, interpretable model (e.g., a linear model or decision tree) locally around the prediction [Ribeiro, Singh, & Guestrin, 2016]. It perturbs the input instance, gets predictions from the black-box model, and then weighs these perturbed samples by their proximity to the original instance to train the local interpretable model.
- **SHAP (SHapley Additive exPlanations):** Based on cooperative game theory, SHAP assigns each feature an importance value (Shapley value) representing its marginal contribution to the prediction, averaged over all possible orderings of features [Lundberg & Lee, 2017]. SHAP values offer desirable properties like local accuracy and consistency.

- **Integrated Gradients (IG):** This method attributes the prediction of a deep network to its input features by integrating gradients along the straight-line path from a baseline input to the actual input [Sundararajan, Taly, & Yan, 2017]. It satisfies axioms like sensitivity and implementation invariance.

3.3.2 Model Distillation and Surrogate Models

These approaches involve training a simpler, more interpretable model (the "student" or "surrogate") to mimic the behavior of a complex black-box model (the "teacher").

- If the surrogate model achieves high fidelity in approximating the black-box model, then understanding the surrogate can provide insights into the original model. Decision trees, linear models, or rule-based systems are often used as surrogates.
- This is a form of global explanation, attempting to capture the overall logic of the complex model.

3.3.3 Concept-based Explanations

Instead of focusing on low-level features (like pixels), these methods aim to explain model decisions in terms of higher-level, human-understandable concepts.

- **TCAV (Testing with Concept Activation Vectors):** TCAV quantifies the importance of user-defined concepts (e.g., "stripes" for zebra classification) for a model's classification decision by examining the directional derivatives of activation vectors in a neural network layer with respect to vectors representing these concepts [Kim et al., 2018].

3.3.4 Example-based Explanations

These methods explain a model's prediction by identifying influential examples from the training data or by generating new examples.

- **Prototypes and Criticisms:** These methods identify representative examples (prototypes) from the training set that are characteristic of a class, and criticisms, which are examples that are poorly represented by the prototypes [Kim, Rudin, & Shah, 2014].
- **Counterfactual Explanations:** These identify the smallest change to an input instance that would alter the model's prediction to a different outcome [Wachter, Mittelstadt, & Russell, 2017]. For example, "Your loan application was denied. If your income had been $5,000 higher, it would have been approved."

3.3.5 Intrinsically Interpretable Models

This approach focuses on designing models that are inherently transparent, rather than trying to explain a black box post-hoc.

- Examples include linear regression, logistic regression, decision trees, rule-based systems, and generalized additive models.
- There is often a trade-off between model performance and interpretability; intrinsically interpretable models may not achieve the same state-of-the-art performance as complex black-box models on very complex tasks. However, research is ongoing to bridge this gap.

3.4 XAI's Connection to Neuroscience: A Two-Way Street

The relationship between XAI and neuroscience is becoming increasingly symbiotic, moving beyond the historical inspiration of AI by neuroscience.

1. **Neuroscience as Inspiration for XAI Paradigms:** How humans explain their own reasoning and decision-making can inform the design of XAI systems. Cognitive science research into human explanation, causality, and mental models can provide valuable principles for developing AI explanations that are truly intelligible and useful to people. For instance, the human tendency to use contrastive explanations ("Why P rather than Q?") is an active area of XAI research.
2. **XAI as a Tool to Understand ANNs (which are Models of the Brain):** This is the most direct and rapidly growing connection. As deep neural networks are increasingly used as predictive models of neural processing in various brain regions (e.g., visual cortex, auditory cortex), XAI techniques become essential tools for neuroscientists. By applying XAI to these ANNs, researchers can:
    - Identify which input features (e.g., visual stimuli characteristics) drive the activity of specific artificial neurons or layers.
    - Compare the "receptive fields" or "tuning properties" of artificial neurons with those of biological neurons.
    - Understand the computational transformations occurring at different stages of the network hierarchy.
    - Generate hypotheses about neural representations and computations that can then be tested experimentally in biological brains.
3. **Evaluating the "Neural Plausibility" of XAI Methods:** Some XAI methods might themselves have more "neurally plausible" interpretations than others. For example, attention mechanisms in ANNs, which highlight relevant parts of an input, bear a conceptual resemblance to attentional processes in the brain and can be considered a form of built-in explainability. Understanding if and how the brain implements processes analogous to XAI techniques could provide deeper insights into both fields.
4. **Bridging Levels of Analysis:** Neuroscience grapples with understanding the brain at multiple levels (molecular, cellular, systems, cognitive). AI models, particularly ANNs, offer a computational bridge between stimulus/behavior (cognitive level) and mechanistic implementation (systems/cellular level). XAI helps to articulate what computations are being performed by the "system" (the ANN) and how its "cellular" components (artificial neurons and their connections) contribute to these computations.

In essence, XAI provides the crucial methodological toolkit for transforming ANNs from mere black-box predictors of neural activity or behavior into interpretable, hypothesis-generating models that can advance our understanding of neural information processing. This makes XAI an indispensable component in the journey towards neural explanation.

4 XAI for Neuroscience: Unveiling the Brain's Mechanisms

The application of Explainable AI (XAI) techniques to neuroscience, particularly in conjunction with deep neural network (DNN) models of brain function, represents a paradigm shift. It moves beyond simply correlating AI model activity with brain activity towards understanding *why* these correlations exist and what computational principles they reveal. This section explores the current landscape of XAI in neuroscience, its successes, and the significant hurdles it faces.

4.1 Current State of the Art: Applications and Insights

XAI methods are being applied across various subdomains of neuroscience, primarily where DNNs have shown promise as functional models of neural systems.

4.1.1 Modeling and Understanding Sensory Processing

The visual system has been a fertile ground for applying XAI to neuro-inspired DNNs.

- **Feature Visualization and Attribution in Vision Models:** CNNs trained on object recognition tasks (e.g., ImageNet) develop hierarchical representations that strikingly resemble those in the primate ventral visual stream [Yamins et al., 2014; Cadieu et al., 2014]. XAI techniques are used to:
  - **Visualize Preferred Stimuli:** Techniques like activation maximization or "deep dream" synthesize input images that maximally activate specific neurons or layers, revealing their "preferred features" [Erhan et al., 2009; Mordvintsev, Olah, & Tyka, 2015]. These can be compared to receptive fields mapped in biological neurons.
  - **Saliency Maps and Attribution:** Methods like Grad-CAM [Selvaraju et al., 2017] highlight image regions most influential for a CNN's classification. These can be compared to human attentional deployment or fMRI activation patterns.
  - **Probing Layer-wise Representations:** By analyzing activations and applying XAI across different layers of a CNN, researchers can map the transformation of information from simple features (edges, textures) in early layers to complex object parts and whole objects in later layers, mirroring the hierarchical processing in the visual cortex [Zeiler & Fergus, 2014].
- **Auditory Cortex Modeling:** Similar approaches are being applied to model the auditory system. DNNs trained on speech recognition or sound event detection learn spectro-temporal features. XAI can help identify which acoustic features drive responses in different model layers, providing hypotheses about auditory feature extraction in the brain [Kell et al., 2018].
- **Olfactory and Somatosensory Systems:** While less developed, efforts are underway to use ANNs and XAI to model other sensory modalities, seeking to understand how complex stimuli are represented and processed.

4.1.2 Decoding Neural Representations and Cognitive States

XAI is not only used to understand models of the brain but also to interpret models that decode brain activity itself.

- **Interpreting Brain Decoders:** Machine learning models are often trained to decode cognitive states (e.g., perceived stimulus, intended movement, emotional state) from neural recordings (fMRI, EEG, ECoG, spikes). XAI techniques can reveal which neural features or brain regions are most informative for these decoding tasks, providing insights into the neural basis of these states [Haufe et al., 2014].
- **Relating ANN Representations to Brain Representations:** Representational Similarity Analysis (RSA) [Kriegeskorte, Mur, & Bandettini, 2008] is a common technique to compare activity patterns in ANNs with those in the brain. XAI can complement RSA by explaining *why* certain ANN layers show high similarity to specific brain regions, by revealing the features that drive those representations.

4.1.3 Understanding Higher Cognitive Functions

The application of XAI to models of more complex cognitive functions is an emerging and challenging area.

- **Decision-Making and Reinforcement Learning:** Reinforcement learning (RL) agents, often implemented with deep neural networks, learn complex policies. XAI methods are being developed to understand the strategies learned by these agents. When these RL agents are used to model

animal or human decision-making, XAI can help elucidate the computational basis of choices and learning processes, potentially linking them to neural circuits like the basal ganglia and prefrontal cortex [Wang et al., 2018, for ANNs in cognitive tasks].

- **Language Processing:** Transformer models have revolutionized NLP. While their primary application is not (yet) as direct models of brain language processing in the same way CNNs are for vision, XAI techniques like attention visualization and feature attribution are used to understand how these models process language [Vig, 2019]. This can generate hypotheses about linguistic representation and computation that may inspire neuroscientific investigation.

4.1.4 Generating Testable Hypotheses

A key promise of XAI in neuroscience is its ability to move beyond descriptive modeling to hypothesis generation.

- By revealing the "computational strategies" or "critical features" used by an ANN that successfully models a brain function, XAI can suggest specific experiments. For example, if XAI indicates a particular texture cue is critical for an ANN's object recognition performance (and the ANN accurately predicts neural responses), neuroscientists can design experiments to test if that specific texture cue is indeed critical for neurons in the corresponding brain area or for behavioral performance.

4.2 Challenges and Limitations

Despite the exciting progress, applying XAI to neuroscience is fraught with significant challenges and limitations.

4.2.1 The "Alignment" or "Correspondence" Problem

This is arguably the most fundamental challenge.

- **Correlation vs. Causation vs. Mechanism:** An ANN might accurately predict neural activity or behavior, and XAI might provide an "explanation" for the ANN's behavior. However, this does not guarantee that the ANN's internal mechanism is the same as the brain's. The ANN could be exploiting statistical regularities in the data or task in a way that is different from the biological system. XAI explains the model, not necessarily the brain directly.
- **Degrees of Freedom and Model Degeneracy:** Many different ANN architectures or training regimens can achieve similar performance on a task or similar correlation with brain activity. XAI applied to these different models might yield different "explanations," making it hard to pinpoint the true neural mechanism.
- **Level of Abstraction:** ANNs are highly abstract models of neural systems. They typically lack biophysical realism (e.g., detailed neuron models, neuromodulation, glial cells, developmental trajectories). Explanations derived from such abstract models might miss crucial biological details or offer misleadingly simple accounts.

4.2.2 Interpretability of the Explanations Themselves

XAI methods produce outputs (e.g., saliency maps, feature importance scores, decision trees). However, these outputs themselves require interpretation by neuroscientists.

- **Complexity of XAI Outputs:** Some XAI outputs can be as complex and hard to understand as the original model, especially for very deep or recurrent networks.

- **Lack of Ground Truth for Explanations:** It's difficult to validate whether an XAI-derived explanation for why an ANN matches brain data is "correct" in a neuroscientific sense. What is the ground truth for a neural computation explanation?
- **Subjectivity and User Bias:** The choice of XAI method, its parameters, and the interpretation of its results can be influenced by the researcher's own biases and hypotheses.

4.2.3 Technical and Methodological Limitations of XAI

XAI techniques themselves have limitations:

- **Fidelity vs. Interpretability Trade-off:** Simpler explanations (e.g., from surrogate models) might be easier to understand but less faithful to the complex model. Highly faithful explanations might be too complex.
- **Robustness and Stability:** Some XAI methods, particularly gradient-based attribution methods, can be sensitive to small input perturbations or model retraining, leading to unstable explanations [Adebayo et al., 2018].
- **Focus on Local Explanations:** Many XAI methods provide local explanations (for a single input) rather than global explanations of the model's overall logic, which might be more relevant for understanding general neural principles.
- **Computational Cost:** Applying sophisticated XAI techniques to very large ANNs and extensive neuroscience datasets can be computationally expensive.

4.2.4 Data Limitations in Neuroscience

The quality and quantity of neuroscience data impose constraints:

- **Data Scarcity and Noise:** Compared to datasets used to train large ANNs (e.g., millions of images), neuroscience datasets are often much smaller, sparser, and noisier. This can limit the complexity of ANN models that can be reliably fit to neural data and, consequently, the insights derivable via XAI.
- **Measurement Limitations:** Current neurotechnologies only capture a small fraction of neural activity, often with limited spatial or temporal resolution. Models built on such partial data might miss crucial aspects of neural computation.

4.2.5 Bridging Different Levels of Analysis

The brain is understood at multiple levels (Marr's levels: computational theory, algorithm/representation, implementation).

- Current ANNs and XAI primarily address the algorithmic/representational level. Connecting these insights to the detailed biophysical implementation in the brain or to the overarching computational theory (the "why" of the computation) remains a major challenge.
- Explanations at the level of artificial neuron activations need careful translation to be meaningful at the level of biological neuron biophysics or circuit dynamics.

4.3 Ethical Considerations

The use of AI and XAI in neuroscience, while promising, also brings forth important ethical considerations, particularly as these insights might eventually translate to applications affecting human health and cognition.

4.3.1 Misinterpretation and Overconfidence

- **False Sense of Understanding:** XAI might provide explanations that appear plausible but are actually misleading or incomplete reflections of true neural mechanisms. This could lead to premature conclusions or misdirected research efforts.
- **Over-reliance on Models:** There's a risk of over-relying on ANN models and their XAI-generated explanations, potentially downplaying the need for direct neurobiological experimentation or alternative theoretical frameworks.

4.3.2 Bias in AI Models and Explanations

- **Propagation of Data Biases:** If ANNs are trained on biased neuroscience data (e.g., from specific demographics, or under particular experimental conditions), they will learn these biases. XAI might then "explain" these biased representations as if they were general principles of brain function.
- **Algorithmic Bias in XAI:** XAI methods themselves might have inherent biases in how they attribute importance or construct explanations, potentially highlighting certain types of features or model components over others.

4.3.3 Neurosecurity and Dual-Use Concerns

- **Manipulation of Neural Systems:** As our understanding of neural computation (aided by AI/XAI) deepens, so too does the potential for technologies that could manipulate brain activity or cognitive states. While potentially beneficial for treating neurological disorders, this also raises concerns about misuse (e.g., for cognitive enhancement in non-medical contexts, or for more invasive forms of influence).
- **Decoding Private Mental States:** If AI/XAI significantly improves our ability to decode thoughts, intentions, or emotions from brain activity, this raises profound privacy concerns.

4.3.4 Clinical Applications and Responsibility

- **Diagnostic and Prognostic Tools:** If AI/XAI-driven insights lead to new diagnostic or prognostic tools for neurological or psychiatric disorders, ensuring their accuracy, fairness, and appropriate use is paramount. Misleading explanations could lead to incorrect diagnoses or treatments.
- **Accountability:** Who is responsible if an AI system, "explained" by XAI, contributes to a negative outcome in a clinical or research setting? The model developers, XAI developers, or the neuroscientists using these tools?

4.3.5 Impact on Scientific Practice

- **Narrowing of Research Focus:** The success of certain AI architectures (e.g., deep learning) might inadvertently narrow the scope of theoretical neuroscience, focusing too much on approaches that fit well with current AI paradigms.
- **Accessibility and Resource Disparities:** The computational resources and expertise required for cutting-edge AI and XAI research can create disparities, potentially disadvantaging researchers or institutions with fewer resources.

Addressing these challenges and ethical considerations requires a cautious, critical, and interdisciplinary approach, involving not only AI researchers and neuroscientists but also ethicists, philosophers, and policymakers. Continuous validation, transparency about limitations, and a commitment to responsible innovation are essential as we navigate this exciting but complex frontier.

5 The Future Trajectory: Towards Deeper Neural Explanation

The symbiotic journey of AI and neuroscience, supercharged by the capabilities of XAI, is poised for even greater integration and impact. The future will likely see a concerted effort to overcome current limitations and develop more sophisticated, biologically-grounded, and interpretable models of the brain. This progression towards deeper neural explanation will require advances on multiple fronts.

5.1 Tighter Integration of AI Model Development with Neuroscientific Constraints

Future AI models aiming to explain the brain will need to incorporate more biological realism, moving beyond purely performance-driven architectures.

- **Biologically Plausible Architectures:** Designing ANNs whose architectures more closely mirror known anatomical structures and connectivity patterns in the brain (e.g., specific cell types, laminar structures, long-range projection systems). This includes exploring more diverse neural units beyond simple point neurons, incorporating dendritic computation, and more realistic synaptic plasticity rules.
- **Incorporating Developmental and Learning Trajectories:** The brain doesn't emerge fully formed; it develops and learns over time through interaction with the environment. Future models might incorporate developmental algorithms and more realistic learning rules (e.g., local, unsupervised, or self-supervised learning that better reflect biological constraints than end-to-end backpropagation alone).
- **Energy Efficiency and Metabolic Constraints:** The brain is remarkably energy-efficient. Incorporating energy constraints into AI model design could lead to sparser, more efficient representations and processing strategies that are more brain-like [Yang et al., 2019, on principles for brain-like computation].
- **Modeling Neuromodulation and Brain States:** Integrating mechanisms analogous to neuromodulatory systems (e.g., dopamine, serotonin, acetylcholine) that can dynamically alter network processing and account for different brain states (e.g., attention, arousal).

5.2 Advancements in XAI for Neuroscience

XAI methods themselves will need to evolve to better suit the specific needs of neuroscientific inquiry.

- **Causal XAI Methods:** Moving beyond correlational feature attribution to methods that can infer causal relationships within the model and, by extension, generate causal hypotheses about the brain. This might involve interventional XAI (e.g., "what if this connection was lesioned?") or counterfactual reasoning at a more mechanistic level.
- **"Neuroscience-Native" XAI:** Developing XAI techniques that are specifically designed to interrogate models in ways that are meaningful to neuroscientists, providing explanations in terms of neural codes, population dynamics, or circuit motifs rather than generic feature importance scores.
- **Explaining Temporal Dynamics:** Many brain functions unfold over time. XAI methods need to become more adept at explaining recurrent neural networks and other models that process temporal sequences, providing insights into dynamic computations.
- **Uncertainty Quantification in Explanations:** Explanations should ideally come with confidence scores or uncertainty estimates, helping researchers gauge the reliability of the insights derived.
- **Interactive and Comparative XAI Tools:** Tools that allow neuroscientists to interactively probe models, compare explanations from different models or XAI methods, and integrate explanations with experimental data visualizations.

5.3 Multi-Modal Data Integration and Modeling

The brain integrates information from multiple senses and controls complex behaviors. Future efforts will increasingly focus on:

- **Modeling Multi-Sensory Integration and Cross-Modal Learning:** Developing ANNs that can learn from and integrate information across different modalities (vision, audition, touch, etc.) and using XAI to understand how these integrated representations are formed.
- **Embodied AI and Sensorimotor Loops:** Using AI agents situated in realistic virtual environments (embodied AI) to model how the brain learns through active interaction with the world. XAI can then be used to understand the learned policies and representations underlying perception-action cycles.
- **Integrating Diverse Neuroscience Data Types:** Building models that can simultaneously account for different types of neural data (e.g., fMRI, ECoG, calcium imaging, single-unit electrophysiology, connectomics) and behavioral data. XAI would then need to synthesize explanations across these data sources.

5.4 Establishing Benchmarks and Best Practices

To ensure rigor and reproducibility, the field will benefit from:

- **Standardized Benchmarking for Brain Models:** Developing common datasets, tasks, and evaluation metrics for assessing how well AI models capture specific brain functions, similar to benchmarks in core AI (e.g., ImageNet, GLUE). Brain-Score is an example of such an effort for vision [Schrimpf et al., 2018; 2020].
- **Best Practices for XAI Application in Neuroscience:** Establishing guidelines for the appropriate use and interpretation of XAI techniques in neuroscientific contexts, including how to report findings, acknowledge limitations, and avoid over-interpretation.
- **Open Science and Collaborative Platforms:** Promoting the sharing of models, data, and XAI tools to facilitate collaboration, replication, and the collective advancement of knowledge.

5.5 Addressing the "Why" Question: Normative and Theoretical Frameworks

While current AI/XAI approaches are powerful for understanding "how" the brain might compute, a deeper understanding also requires addressing "why" the brain computes in a particular way.

- **Connecting with Normative Theories:** Linking AI models and their XAI-derived explanations to normative theories from computational neuroscience and cognitive science (e.g., Bayesian inference, predictive coding, efficient coding). Does the model's strategy approximate an optimal solution to a computational problem faced by the organism?
- **Evolutionary and Developmental Perspectives:** Considering how evolutionary pressures and developmental processes might have shaped neural architectures and computational strategies.

5.6 Continued Dialogue on Ethical Implications

As the capabilities for neural explanation grow, ongoing ethical reflection will be crucial.

- **Proactive Ethical Frameworks:** Developing ethical guidelines and governance structures proactively, rather than reactively, to address the societal implications of advanced brain understanding and potential brain-interfacing technologies.
- **Public Engagement and Education:** Fostering public understanding of the goals, capabilities, and limitations of AI in neuroscience to ensure informed societal discourse.

The future trajectory is one of increasing sophistication and integration. The goal is not just to build AI that mimics brain-like intelligence but to leverage AI and XAI as indispensable scientific instruments to achieve a principled, mechanistic, and interpretable understanding of biological intelligence itself. This endeavor promises to be one of the most exciting and transformative scientific frontiers of the 21st century.

6 Conclusion

The intertwined history of Artificial Intelligence and Neuroscience is a testament to the power of interdisciplinary cross-pollination. From the earliest days when the neuron served as a blueprint for computational thought, to the current era where complex AI models are becoming sophisticated tools for dissecting neural circuits, their relationship has evolved into a deeply symbiotic one. This journey has not been linear; it has seen periods of fervent collaboration, followed by divergence as each field pursued its own specialized goals, only to re-converge with renewed vigor, driven by new theoretical insights and technological breakthroughs.

The advent of deep learning has heralded a particularly exciting phase in this partnership. AI systems, initially designed for engineering purposes, now offer unprecedented opportunities to model and understand the brain's information processing capabilities. However, the inherent "black box" nature of these powerful models presented a significant barrier to their scientific utility for neuroscience. This is where Explainable AI (XAI) has emerged as a transformative force. By providing tools to peer inside these complex AI systems, XAI offers a critical bridge, enabling researchers to translate the internal workings of AI models into testable hypotheses about neural mechanisms. The ambition has shifted from mere neural inspiration for AI, to AI facilitating genuine neural explanation.

We have explored how XAI techniques are currently being applied to analyze AI models of sensory systems, decode neural representations, and even probe higher cognitive functions. These applications are beginning to yield novel insights and generate specific, falsifiable predictions about how the brain works. Yet, the path towards comprehensive neural explanation is laden with challenges. The "alignment problem"—ensuring that explanations of AI models accurately reflect biological reality—remains paramount. The interpretability of XAI outputs themselves, the technical limitations of current XAI methods, the constraints of neuroscience data, and the difficulty of bridging different levels of biological analysis all demand careful consideration and innovative solutions.

Furthermore, the ethical dimensions of this endeavor cannot be overstated. As AI provides deeper insights into the brain, the potential for both benefit and misuse grows. Concerns regarding misinterpretation, bias, neurosecurity, and the responsible development of clinical applications necessitate ongoing vigilance, robust ethical frameworks, and broad societal dialogue.

Looking ahead, the future of AI and neuroscience synergy, powered by XAI, promises even deeper integration. We anticipate AI models that are more biologically constrained, XAI methods that are more causally informative and tailored to neuroscientific questions, and a greater emphasis on multimodal data integration and normative understanding. The development of standardized benchmarks and collaborative platforms will be crucial for accelerating progress and ensuring rigor.

In conclusion, the journey from neural inspiration to neural explanation is reshaping our understanding of intelligence in both its artificial and biological manifestations. It is a journey that requires humility in the face of the brain's immense complexity, creativity in developing new tools and theories, and a steadfast commitment to ethical principles. While the ultimate goal of a complete understanding of the brain may still be distant, the symbiotic relationship between AI and neuroscience, fortified by the principles of explainability, offers one of the most promising paths toward unraveling the mysteries of the mind. This partnership is not just advancing two distinct scientific fields; it is forging a new, integrated science of intelligence.

References (Illustrative Examples)

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.
- Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., ... & DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS computational biology, 10*(12), e1003963.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition (CVPR)*.
- Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal, 1341*(3), 1.
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage, 87*, 96-110.
- Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*. Wiley.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735-1780.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences, 79*(8), 2554-2558.
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology, 160*(1), 106.
- Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron, 98*(3), 630-644.
- Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology, 10*(11), e1003915.
- Kim, B., Rudin, C., & Shah, J. (2014). The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *International conference on machine learning (ICML)*.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological cybernetics, 43*(1), 59-69.
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience, 2*, 4.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems (NeurIPS)*.

- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278-2324.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems (NeurIPS)*.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT press.
- McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2: Psychological and biological models*. MIT press.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133.
- Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press.
- Mordvintsev, A., Olah, C., & Tyka, M. (2015). Inceptionism: Going deeper into neural networks. *Google AI Blog*.
- Ramón y Cajal, S. (1894). The fine structure of the nervous system. *In Croonian Lecture*. (Translation available).
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience, 2*(11), 1019-1025.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review, 65*(6), 386.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533-536.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., ... & DiCarlo, J. J. (2018). Brain-Score: Which Artificial Neural Network for Vision Best Resembles the Human Brain?. *bioRxiv, 407007*. (Later published in PNAS 2020)
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593-1599.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision (ICCV)*.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*(7587), 484-489.
- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning (ICML)*.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems (NeurIPS)*.
- Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. *Proceedings of the 2019 ACL System Demonstrations*.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech., 31*, 841.
- Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience, 21*(6), 860-868. (This is an example of ANNs in cognitive tasks, not XAI of RL agents specifically for neuroscience, but relevant context).
- Werbos, P. J. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences* (Doctoral dissertation, Harvard University).
- Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *1960 IRE WESCON Convention Record, Part 4*, 96-104.
- Wiener, N. (1948). *Cybernetics: Or control and communication in the animal and the machine*. MIT Press.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*(23), 8619-8624.
- Yang, G.R., Joglekar, M.R., Song, H.F., Newsome, W.T., Wang, X.J. (2019). Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience, 22*(2), 297-306. (Illustrates principles of computation).
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision (ECCV)*.