information

MDPI

Review

The Black Box Paradox: AI Models and the Epistemological Crisis in Motor Control Research

Nuno Dias [1,2], Liliana Pinho [1,2], Sandra Silva [1,2], Marta Freitas [1,2], Vânia Figueira [1,2] and Francisco Pinho [1,2,*]

[1] Escola Superior de Saúde do Vale do Ave, Cooperativa de Ensino Superior Politécnico e Universitário, Rua José António Vidal, 81, 4760-409 Vila Nova de Famalicão, Portugal; nuno.dias@ipsn.cespu.pt (N.D.); liliana.pinho@ipsn.cespu.pt (L.P.); sandra.silva@ipsn.cespu.pt (S.S.); marta.goncalves@ipsn.cespu.pt (M.F.); vania.figueira@ipsn.cespu.pt (V.F.)

[2] HM—Health and Human Movement Unit, Cooperativa de Ensino Superior Politécnico e Universitário, CRL, 4760-409 Vila Nova de Famalicão, Portugal

* Correspondence: francisco.pinho@ipsn.cespu.pt

Abstract

The widespread adoption of deep learning (DL) models in neuroscience research has introduced a fundamental epistemological paradox: while these models demonstrate remarkable performance in pattern recognition and prediction tasks, their inherent opacity contradicts neuroscience's foundational goal of understanding biological mechanisms. This review article examines the growing trend of using DL models to interpret neural dynamics and extract insights about brain function, arguing that the black box nature of these models fundamentally undermines their utility for mechanistic understanding. We explore the distinction between computational performance and scientific explanation, analyze the limitations of current interpretability techniques, and discuss the implications for neuroscience research methodology. We propose that the field must critically evaluate whether DL models can genuinely contribute to our understanding of neural processes or whether they merely provide sophisticated curve-fitting tools that obscure rather than illuminate the underlying biology.

Keywords: artificial intelligence; deep learning; motor control; computational neuroscience; interpretability

check for updates

Academic Editor: Birgitta Dresp-Langley

Received: 29 July 2025
Revised: 21 September 2025
Accepted: 23 September 2025
Published: 24 September 2025

Citation: Dias, N.; Pinho, L.; Silva, S.; Freitas, M.; Figueira, V.; Pinho, F. The Black Box Paradox: AI Models and the Epistemological Crisis in Motor Control Research. *Information* **2025**, *16*, 823. https://doi.org/10.3390/info16100823

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Neuroscience stands at a crossroads. The field has witnessed an unprecedented influx of Artificial Intelligence (AI) methodologies, driven by their remarkable success in several applications. From analyzing neural recordings to modeling brain networks, machine learning (ML) and deep learning (DL) approaches have become increasingly prevalent in neuroscience research [1–4]. However, this shift raises a fundamental question: can models that operate as black boxes truly advance our understanding of biological systems that we seek to comprehend mechanistically [5,6]? The allure of DL in neuroscience is understandable. These models can process vast amounts of neural data, identify complex patterns, and make accurate predictions about neural activity. They promise to unlock insights from high-dimensional datasets that traditional analytical methods struggle to handle [7–9]. Yet, beneath this computational prowess lies a profound epistemological challenge that the field has yet to address adequately. The central thesis of this review is that the opaque nature of AI models creates an irreconcilable tension with some of

neuroscience's fundamental objectives, particularly in motor control research. While these models excel at pattern recognition and prediction, they are also vulnerable to functioning as little more than sophisticated curve-fitting tools [10], and their opacity prevents the mechanistic understanding that is essential for advancing our knowledge of brain function. This paradox becomes particularly problematic when researchers attempt to extract biological insights from AI models trained on neural data, essentially trying to understand one complex system (the brain) through another complex system (the artificial neural network (ANN)) whose internal workings remain largely cryptic.

It is important to acknowledge, however, that some subfields of neuroscience, particularly cognitive neuroscience, have already engaged critically with the epistemic limits of deep learning. For instance, Kanwisher et al. [11] explicitly caution against mechanistic overreach in ANN-based modeling, highlighting the risks of misinterpreting predictive performance as mechanistic insight. Recognizing these ongoing debates situates our argument within a broader critical discourse, while underscoring that motor control research has not yet fully incorporated such cautionary perspectives. While the distinction between prediction and explanation in deep learning has been widely discussed across AI, philosophy of science, and cognitive neuroscience, the novelty of this paper lies in applying this critique specifically to the field of motor control. We argue that DL adoption in this domain, which is already conceptually fragmented, risks exacerbating existing theoretical uncertainties if not done within a deeply comprehensive understanding of these tools, their characteristics and consequently, their limitations. Thus, the contribution of this manuscript is field-specific: to demonstrate how reliance on opaque models compounds epistemological fragmentation in motor control research, rather than to generalize a critique to all neuroscience. Our critique is specifically aimed at the use of DL for mechanistic inference in motor control, rather than at its well-recognized utility for engineering-style prediction.

The paper proceeds as follows. First, we examine the black box problem, outlining the intrinsic opacity of DL models and the limitations of current interpretability techniques (Section 2). We then explore the implications of these issues in motor control research a field whose conceptual fragmentation makes it particularly vulnerable to misinterpretations of predictive models (Section 3). Next we situate these concerns within broader epistemological debates about explanation, causality, and scientific understanding (Sections 4 and 5). Finally, we consider constructive directions for future research, including revised evaluation criteria and hybrid modeling approaches that embed mechanistic priors into DL architecture (Section 6).

2. The Nature of DL Black Boxes

2.1. Defining the Black Box Problem

The characterization of AI models as "black boxes" is not merely a metaphorical description but a fundamental property that emerges from their architecture and learning mechanisms [12–14]. Unlike traditional statistical models, where parameters generally have clear interpretations and relationships can be explicitly defined, DL models operate through layers of nonlinear transformations that create emergent computational behaviors. In a typical deep neural network, information flows through multiple hidden layers, each applying weighted transformations and nonlinear activation functions [15–17]. The final output emerges from this cascade of operations, but the intermediate representations and their contributions to the final result remain largely undefined. Even when we can examine the weights and activations within the network, their high-dimensional nature and complex interactions make meaningful interpretation extremely challenging [18–20]. This opacity is not an accidental byproduct of DL but rather an inherent consequence of its design. DL models are optimized for performance, not interpretability [18,19]. The learning algorithms

adjust millions or billions of parameters to minimize loss functions, creating internal representations that may be effective for the task at hand but might bear no necessary relationship to the underlying biological processes they are meant to model [14,21,22].

2.2. The Performance-Interpretability Trade-Off

The success of DL has been built on a fundamental trade-off: sacrificing interpretability for performance. This trade-off is explicit and intentional—or, at least, justifiable—in most AI applications [23,24]. When developing a computer vision system or natural language processor, the primary concern is accuracy, not understanding how the system works. The black box nature is acceptable because the goal is functional performance, not mechanistic insight.

However, this trade-off becomes problematic when DL models are applied to scientific domains where understanding mechanisms is the primary objective. In neuroscience, we are not merely interested in predicting neural activity or even functional outcomes. We want to understand how neural circuits compute, how different brain regions interact, and how these processes give rise to behavior. This mechanistic understanding is not merely an academic pursuit but a critical foundation for multiple research and clinical domains [25–27]. Motor control research requires deeper insight into movement orchestration mechanisms to develop effective rehabilitation frameworks for conditions of movement dysfunction. Cognitive scientists need to understand the building blocks of neural function to advance psychological/psychiatric rehabilitation practices. The performance-interpretability trade-off that works well for engineering applications fundamentally conflicts with these scientific goals.

2.3. Limitations of Current Interpretability Techniques

The AI community has developed various techniques to peer inside the black box, including gradient-based attribution methods [28], attention mechanisms [29,30] and layer-wise relevance propagation [31]. While these methods provide some insights into model behavior, they may fall short of providing the mechanistic understanding required for scientific inference. Gradient-based methods, for instance, can identify which input features most strongly influence the output, but they cannot explain why these features are important or how they interact within the model [28]. Attention mechanisms show where the model 'focuses' its computation, but attention weights do not necessarily correspond to causal relationships or meaningful biological processes [29,30]. LRP methods face fundamental interpretability limitations due to their dependency on architectural choices, parameter selection, and reference point definitions, requiring extensive manual tuning and subjective validation while lacking objective quality metrics—ultimately making explanations as much about methodological decisions as about actual model behavior [31]. More fundamentally, these interpretability techniques often provide post hoc explanations that may not reflect the actual computational processes within the model. They are interpretations of interpretations, adding additional layers of uncertainty to an already opaque system.

3. The Motor Control Conundrum: Why AI Models Might Exacerbate Foundational Uncertainties

3.1. The Unresolved Theoretical Landscape

Motor control research finds itself in a peculiar epistemological position: despite decades of sophisticated experimentation using neural recordings, electromyographic data, kinematic analysis, and computational modeling, the field remains fundamentally divided on its most basic conceptual foundations. This fragmentation is not merely academic—it directly impacts our ability to develop effective rehabilitation strategies and understand

movement disorders. Consider the persistent ambiguity surrounding the concept of "motor control" itself. Is it the orchestration of muscle activations to achieve behavioral goals [32,33]? The specification of movement trajectories [34,35]? The management of mechanical redundancy [36]? Or, as Feldman's equilibrium point hypothesis suggests [37], the setting of threshold positions for a proposed parameter of neuromuscular activation? Each definition implies radically different experimental approaches and interpretations of neural data.

3.2. The Parameter-Variable Distinction Crisis

Perhaps no debate better exemplifies the field's theoretical crisis than the unresolved question of what constitutes control parameters versus control variables in movement [36,37]. Traditional approaches often treat observable quantities—joint angles, muscle forces, movement velocities—as the controlled entities. However, this assumption faces serious challenges. First, the computational implausibility: if the nervous system directly controlled all mechanical variables, it would require solving complex inverse dynamics problems in real-time, a computational burden that seems biologically unrealistic given neural processing constraints. Second, the empirical evidence suggests the nervous system may operate at a more abstract level. Neural manifold analyses reveal that motor cortical activity occupies low-dimensional subspaces [9,38], suggesting control might occur through specification of manifold states rather than explicit variable control. Yet, what these manifolds represent—movement goals, muscle synergies, or abstract control laws—remains contentious.

3.3. Competing Frameworks: Irreconcilable or Complementary?

The field's theoretical landscape resembles a dynamic confluence of competing paradigms, each with its own experimental support and explanatory gaps (Computational approaches grounded in optimal control theory propose that the brain maintains internal models that predict sensory consequences of motor commands [39]. These frameworks excel at explaining adaptation and motor learning but struggle to account for the flexibility and context-sensitivity of natural movement. Dynamical systems perspectives reject the notion of explicit internal models, instead viewing movement as emerging from the interaction of neural dynamics [9,40,41], biomechanics, and environmental constraints. While capturing movement's fluidity, these approaches often lack the predictive precision needed for clinical applications. Ecological approaches [42–44] emphasize perception-action coupling and argue that movement control cannot be understood in isolation from environmental affordances. Yet translating these insights into quantifiable neural mechanisms remains elusive. This theoretical plurality presents a fundamental challenge for AI-based approaches: without consensus on the underlying principles of motor control, AI models inevitably embed the biases and limitations of whichever framework they adopt, while simultaneously lacking the capacity to adjudicate between fundamentally different conceptual approaches to biological control.

3.4. The Data-Theory Gap in Motor Control Neuroscience Research

Modern recording techniques generate unprecedented volumes of neural data—hundreds of simultaneously recorded neurons, millisecond-precision spike trains, complex patterns of cortical-spinal interactions. Paradoxically, this data deluge has not resolved fundamental theoretical questions but often deepened them. Consider directional tuning in motor cortex: initially interpreted as evidence for explicit movement direction encoding [45–47], subsequent analyses revealed that apparent tuning could emerge from underlying dynamical processes unrelated to explicit directional coding [35,48]. The same neurons, the same data, yet fundamentally different interpretations—each mathematically valid, neither definitively falsifiable.

3.5. Why DL Might Struggle to Resolve These Debates

In this contested theoretical landscape, DL introduces further representational opacity. The fundamental problem is that DL models trained on motor neural data will inevitably reflect the theoretical assumptions embedded in their training paradigms, while their ambiguous nature prevents us from identifying which assumptions drive their outputs. A concrete example illustrates this dilemma: Suppose we train a deep network to decode intended movements from motor cortical activity. The network achieves impressive accuracy, perhaps even surpassing traditional decoders. What have we learned about motor control? The network might be exploiting motor command signals reflecting the brain's control strategy, incidental correlations between neural activity and movement kinematics, feedback-related signals that correlate with but don't cause movement, preparatory activity that precedes but doesn't directly specify movement or even statistical regularities in the data that have no biological significance. Without mechanistic interpretability, we cannot distinguish among these possibilities. The model's success tells us it has found predictive patterns, but not whether these patterns reflect the brain's actual control principles. An illustrative case is summarized in Box 1.

Box 1. Center-out reaching tasks.

Deep learning decoders often achieve high accuracy when trained on motor cortical activity during center-out reaching tasks. This confirms that population activity contains directionally relevant signals. However, such predictive success does not clarify whether the motor cortex explicitly encodes direction, whether trajectories emerge from dynamical states, or whether apparent tuning curves reflect causal control variables. Without mechanistic interpretability, these results remain correlational rather than explanatory.

It is also crucial to differentiate between DL applications that are purely pragmatic—such as signal decoding for brain–computer interfaces (e.g., LSTM-based cursor control)—and those that implicitly or explicitly claim to reveal internal neural representations [49]. The epistemic implications of these uses differ substantially: the former are primarily engineering solutions, while the latter risk being overinterpreted as mechanistic insights into motor cortical organization. Our critique targets primarily the latter category.

Prominent examples from other domains highlight this issue. In vision research, Yamins & DiCarlo [50] and Kriegeskorte [51] used and reviewed goal-driven DL models to capture properties of sensory cortex, but their work is sometimes misread as providing mechanistic accounts rather than functional correspondences. In motor control, Willett et al. [49] demonstrated high-performance brain-to-text communication from motor cortical signals. While methodologically groundbreaking, this study is sometimes construed as revealing the "neural code for handwriting" when in fact it provides a powerful decoding tool without mechanistic explanation of motor cortical dynamics. These cases exemplify the risk of conflating predictive success with mechanistic understanding.

A consolidated set of examples from motor control further illustrates this epistemological tension. Recurrent neural networks trained to reproduce reaching trajectories and motor cortical dynamics [52] achieve strong predictive alignment with observed activity but do not adjudicate between competing mechanistic frameworks such as optimal feedback control, internal models, dynamical models or equilibrium-point hypotheses nor do they allow for consolidation of such theories. Similarly, sequential auto-encoders for neural population activity [53] yield smoother single-trial trajectories and improved decoding, yet the learned latent factors are not identified with specific mechanistic elements (e.g., control parameters, referent states or prediction-error computations). In brain–computer interfaces, high-performance decoders demonstrate practical power without clarifying the representational or causal computations of motor cortex [49]. Taken together, these cases highlight

how DL can succeed technically while leaving core mechanistic questions unresolved. A classic paradigm, force-field adaptation, is detailed in Box 2 to illustrate this tension.

Box 2. Force-field adaptation.

Force-field paradigms have historically been used to support internal model theories, as they reveal structured aftereffects and generalization patterns. DL models can fit adaptation curves and predict hand trajectories under perturbations, demonstrating strong statistical alignment. Yet this does not prove the existence of forward or inverse models in the brain. Unless prediction-error variables or architectural constraints are built into the model, DL-based success risks being misinterpreted as mechanistic evidence.

This concern has been raised by others who attempt to bridge neuroscience and deep learning. Marblestone, Wayne, and Kording [54] argue that integration between these fields requires not only technical advances but also conceptual clarity about the nature of explanation. Similarly, Barrett, Morcos, and Macke [22] caution against overinterpreting internal representations in ANNs as if they mirrored those in biological systems, highlighting the limits of representational claims in DL.

This risk reflects a broader epistemological tension: while deep learning systems can be optimized for predictive accuracy, this does not guarantee mechanistic insight into the underlying biological system. From a mechanistic perspective [25], explanatory models must specify the organized structural components and their causal interactions. Likewise, interventionist theories of explanation [55] demand that models support counter factual reasoning—i.e., allow us to infer what would happen if a particular component of the system were altered. DL models, unless explicitly constrained by biologically grounded priors, often fail to meet these standards. Our argument builds on these explanatory frameworks to clarify where and how predictive models can—and cannot—contribute to scientific understanding in neuroscience. This refinement ensures that our critique does not overgeneralize but instead targets those instances where predictive models are at risk of being overinterpreted as mechanistic explanations.

3.6. The Clinical Stakes

This is not merely an academic concern. Motor rehabilitation is often informed by understanding what aspects of movement control are impaired in conditions like stroke, Parkinson's disease, spinal cord injury or even musculoskeletal disorders [56]. If we train DL models to predict recovery outcomes or optimize therapeutic interventions without understanding the underlying control principles, we risk developing interventions that target symptoms rather than causes, missing opportunities to leverage the nervous system's actual control mechanisms, creating rehabilitation protocols that work in specific contexts but fail to generalize or perpetuating theoretical confusions that impede scientific progress. The promise of DL in motor control—to find patterns in complex neural data that reveal fundamental principles—thus confronts an immense obstacle: without prior resolution of what we're looking for (parameters vs. variables, computational vs. dynamical control, hierarchical vs. distributed organization), AI models may ultimately reproduce theoretical ambiguities through more complex mappings, rather than resolving them.

At the same time, it is important to acknowledge that DL has already shown clinical value in contexts where predictive accuracy is sufficient for decision-making. For example, DL-based approaches have improved diagnostic performance in medical imaging [57] and prediction of functional outcomes in stroke [58], despite lacking mechanistic transparency. Such applications demonstrate that DL can be clinically useful as a statistical predictor. However, the situation differs fundamentally when therapeutic success depends on causal understanding, as in personalized neurostimulation or targeted neurorehabilitation. In

these cases, explanatory adequacy is indispensable: without insight into how neural circuits generate and control movement, predictive models risk misdirecting interventions. Thus, DL can support clinical practice where prediction suffices but mechanistic clarity remains essential for domains where intervention must act directly on causal neural processes.

3.7. The Path Forward Requires Theoretical Clarity

The notion of an "epistemological crisis" in motor control is not merely rhetorical: it becomes concrete when black-box DL approaches are contrasted with established mechanistic theories of movement control. Each of the major frameworks in motor control research articulates specific explanatory commitments. Feldman's equilibrium-point hypothesis (or λ-model) grounds movement generation in shifts of referent configurations: muscles are activated through the shift of a specified threshold, linking neural commands, biomechanics, and reflex gains [37]. Optimal feedback control specifies explicit cost functions, state estimators, and control policies, predicting characteristic signatures such as the minimum intervention principle and the alignment of variability with task-irrelevant dimensions [59,60]. Internal forward and inverse models assume efference copies and predictive state estimates, explaining phenomena such as sensory attenuation and structured aftereffects during adaptation [40,61].

By contrast, DL models typically optimize statistical mappings from neural activity to behavior without encoding these mechanistic and observable commitments. A high-performing decoder may predict hand trajectories or muscle activity, but it does not test for the existence of referent configurations, cost functions, or prediction-error computations [49–51]. In this sense, DL sidesteps the central theoretical assumptions that have historically structured debates in motor control.

The crisis, therefore, lies in the growing misalignment between methodological power and explanatory adequacy. DL provides powerful predictive tools, but when their success is conflated with a mechanistic explanation, theory testing is deprioritized and long-standing debates risk being bypassed rather than resolved perpetuating the neglect of a much needed logically sound and experimentally verifiable consolidation of knowledge. Before motor control research can benefit fully from advanced computational tools, it must address its foundational uncertainties. This requires experimental paradigms that distinguish between competing theories, the development of formal mathematical bridges between abstract models and neural implementation, and clearer criteria for what counts as a satisfactory explanation of motor control. Only with such theoretical grounding can DL approaches contribute to genuine mechanistic understanding rather than sophisticated pattern recognition.

To make these theoretical contrasts more operational, we synthesized the main motor-control frameworks in Table 1. This comparative summary highlights their core assumptions, key predictions, decisive paradigms, and the extent to which DL approaches can test these commitments.

This table ultimately reveals a striking irony: the same definitional ambiguity that makes these frameworks difficult to empirically distinguish from each other also makes them impossible to properly test with DL. When we can't agree on what constitutes variable of parameter in biological terms, how can we expect an AI-based approach to illuminate these distinctions? The frameworks most amenable to DL investigation—like dynamical systems with its observable neural trajectories—are precisely those making the fewest claims about specific control mechanisms, while those proposing explicit architectures remain opaque to deep investigation. This creates a double bind: we're not just unfit to judge whether DL implements these mechanisms, we're often unfit to judge whether the biological system implements them either. Perhaps this persistent heterogeneity isn't a

failure of scientific progress but a reflection of motor control's nature—the brain might employ bits of optimal control here, something resembling internal models there, threshold control elsewhere, or entirely different principles we haven't conceived. The table thus becomes a proxy to display current limitations, a structured admission that after decades of research, we have multiple incompatible yet partially successful theories we can neither definitively prove nor disprove, and DL might contribute to further obscure rather than illuminate such ambiguity. Rather than attacking DL's application to motor control, this comparison calls for intellectual honesty about what remains unknown and what current methods—biological or artificial—cannot tell us.

Table 1. Comparative summary of motor-control frameworks.

Framework	Core Assumptions	Key Predictions/Empirical Signatures	Decisive Paradigms	Can DL Test This? (How/Limits)
Optimal Feedback Control (OFC)	Goal-directed actions optimized by internal cost functions and state estimation.	Minimal Intervention Principle (MIP); variability aligned with task-irrelevant dimensions; rapid, context-dependent feedback gains.	Mechanical/visual perturbations; obstacle/anisotropy tasks; variance–covariance alignment.	**Partly.** DL might provide insight over motor output and outcomes but does not allow to infer directly over the relevant cost function being minimized.
Internal Models (Forward/Inverse)	Predictive mappings. Sensorimotor loop of motor and sensory outcomes dependent on inverse modeling of motor commands.	After-effects (e.g., prism/force-field washout); context-specific generalization; sensory attenuation.	Prism/curl-field adaptation; visuomotor rotations; context-switching.	**Partly.** DL can capture adaptation curves. Mechanistic inference over command architecture requires prediction-error variables or forward-model priors.
Dynamical Systems	Movement emerges from low-dimensional neural dynamics/manifolds; preparatory activity sets initial conditions of motor output.	Rotational dynamics; low-dimensional manifolds; preparatory-to-movement transitions; robustness to perturbations; often diffuse encoding of motor variables.	Neural population recordings; transient cortical perturbations; dynamical fits vs. tuning models.	**Partly.** DL latent-state models (RNNs, sequential auto-encoders) can recover trajectories and map latent variables. Insight over biological plausibility remains limited.
Equilibrium-Point (λ-model)	CNS specifies the threshold of toni-stretch reflex-based parameter lambda (λ)	EMG patterns from referent shifts; unloading effects; posture-to-movement continuity.	Unloading/force perturbations; tonic stretch reflex manipulations.	**No.** The core parameter and assumption of this model remains theoretical and experimentally undefined beyond logic and interpretation.
Ecological/Perception–Action	Control grounded in affordances; tight perception–action coupling.	No central executive of movement organization; direct perception; self-organization; affordance-based organization.	Constraints-led approach; affordance-based motor outcomes;	**No.** Sensory and motor activity may be traced but core assumptions remain untestable beyond fragile inference of outcomes.

Notes. "Can DL test this?" refers to mechanistic, not merely predictive, tests. Most cases require causal perturbations and explicit priors (e.g., cost functions, referent states, affordance variables). This table complements the evaluation rubric in the table in Section 6.

4. The Epistemological Challenge in Neuroscience

4.1. Scientific Understanding vs. Predictive Performance

This section examines the epistemological challenges posed by AI and DL in neuroscience, contrasting predictive performance with the requirements for mechanistic explanation.

The adoption of AI in neuroscience reflects a broader tension between different conceptions of scientific knowledge. On one hand, there is the pragmatic view that successful prediction constitutes understanding—if a model can accurately predict neural activity, it may appear to have captured something essential about the system, but as philosophers of science note, prediction does not by itself imply explanation [62]. On the other hand, there is the mechanistic view that true understanding requires knowledge of how the system works, not just what it does. This tension is not unique to neuroscience, but it is particularly acute in this field because of the complexity of the brain and our limited ability to directly observe its computational processes beyond the physical and measurable outcomes. Traditional neuroscience has relied heavily on mechanistic reasoning, building understanding through careful experimentation, theoretical modeling and the development of increasingly sophisticated techniques for measuring and manipulating neural activity. AI models, however, operate on a fundamentally different epistemological foundation. They are trained to optimize objective functions without proper regard for biological plausibility or mechanistic interpretability. The representations they learn may be statistically useful for prediction tasks, but they may bear no relationship to the actual computational processes in the brain.

In neuroscience, however, explanation has a more precise meaning than in general philosophy of science. Mechanistic explanation requires specifying the organized entities and activities that give rise to a phenomenon [25,26]. Interventionist causality emphasizes that an adequate explanation should allow predictions about what would happen under counterfactual interventions [55]. Similarly, in computational neuroscience, explanatory models are often expected to respect biophysical interpretability and to align with known neural mechanisms. These perspectives set stricter criteria than predictive accuracy, and it is against such criteria that DL models often fall short.

However, to assess the true explanatory power of DL models in neuroscience, it is essential to clarify what counts as explanation in this domain. In computational neuroscience, mechanistic explanation typically refers to identifying the organized biological entities (e.g., neurons, synapses, populations) and their structured interactions that give rise to specific phenomena [25,26]. According to this framework, a model is explanatory if it mirrors the causal organization of the system it aims to describe.

Similarly, interventionist theories of explanation, such as those proposed by Woodward, assert that an adequate model should support counterfactual reasoning; that is, it should allow us to predict what would happen if a specific component were altered [55]. A further explanatory criterion is biophysical interpretability, which refers to how closely the components and parameters of a model correspond to known neurophysiological mechanisms. Recent hybrid architectures—such as spiking neural networks [63] and biologically plausible deep learning models implementing local synaptic learning rules [64]—have sought to embed biological constraints into DL systems, thereby enhancing their explanatory potential. These developments aim to bridge the gap between high model performance and mechanistic plausibility in neural computation. Despite their promise, such biologically constrained approaches remain largely underutilized in motor control research.

Without adhering to at least one of these explanatory standards—mechanistic, causal, or biophysical—DL models may function primarily as sophisticated statistical tools without contributing to genuine scientific understanding. This distinction is essential for evaluating not only what DL models can predict, but what they can help us explain.

Beyond these explanatory standards, a deeper engagement with explainable AI approaches is also warranted in neuroscience. First, it is important to distinguish between post hoc interpretability techniques (e.g., saliency maps, feature attribution, counterfactual analysis) and intrinsically interpretable models designed with transparency as a goal. Second, interpretability methods differ in epistemic strength: gradient-based saliency provides correlational attributions, symbolic or rule-based methods yield human-readable structures, while causal abstraction methods seek to align learned representations with known causal structures. Third, hybrid approaches—such as constraining DL with domain-specific priors (e.g., sparsity, local dynamics, hierarchical modularity)—offer promising avenues for embedding neuroscientific plausibility directly into architectures. Finally, recent work on causal DL and abstraction alignment [5,12,21] highlights the potential for models that not only predict but also support interventionist reasoning.

These strategies remain underexplored in motor control, but they provide concrete avenues for making DL models more than black-box predictors and for aligning computational advances with neuroscientific standards of explanation.

4.2. The Problem of Multiple Realizability

A fundamental challenge in using AI models to understand neural processes is the problem of multiple realizability—the fact that the same input-output relationship can be implemented by vastly different internal mechanisms [10,65]. This is particularly relevant for DL models, which can achieve similar performance through radically different weight configurations and internal representations [54,65]. Consider a deep neural network trained to classify visual stimuli based on neural recordings from the visual cortex. Even if the model achieves high accuracy, there are countless ways the network could have learned to perform this task, most of which may bear no resemblance to the actual computations performed by biological neural circuits. The model might rely on statistical regularities in the data that are irrelevant to biological processing, or it might develop internal representations that are optimized for the artificial task rather than reflecting genuine neural mechanisms [10,13]. This problem is compounded by the fact that DL models are typically trained on limited datasets that may not capture the full complexity and variability of biological systems. The learned representations may be overfitted to the specific characteristics of the training data rather than reflecting general principles of neural computation.

4.3. The Circularity Problem: Explaining the Brain with Artificial Brains

Another fundamental epistemological issue is the circularity that arises when artificial neural networks (ANNs) are used to explain biological neural networks. When researchers train DL models on neural data and then attempt to extract insights about brain function from these models, they are essentially trying to understand one complex, high-dimensional system through another complex, high-dimensional system. This approach assumes that ANNs, despite their differences from biological networks, can somehow reveal truths about biological computation [11]. However, this assumption is questionable on several grounds. First, ANNs are based on highly simplified models of biological neurons that ignore many important aspects of neural computation, including temporal dynamics, neuromodulation, and the complex morphology of real neurons. Second, the learning algorithms used to train artificial networks bear little resemblance to the learning mechanisms in biological networks, which are themselves not profoundly clear. The optimization landscapes, training procedures, and architectural constraints are fundamentally different between artificial and biological systems [66,67]. The conceptual space of prediction versus explanation is summarized in Figure 1, which situates DL and mechanistic approaches along axes of predictive accuracy and mechanistic insight.

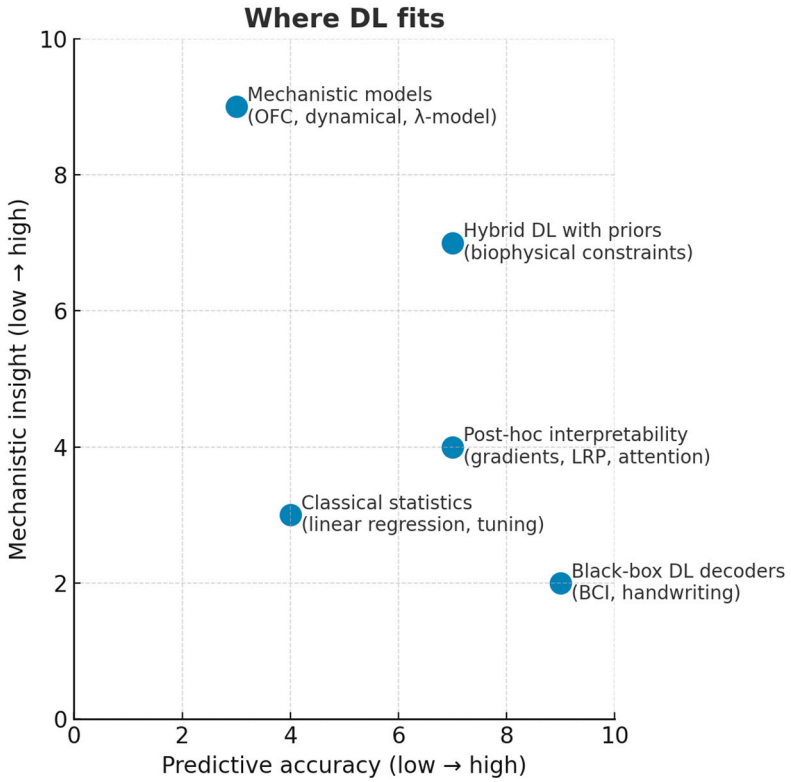


Figure 1. Where DL fits. Conceptual positioning of different modeling approaches in motor control and neuroscience along two axes: predictive accuracy (*x*-axis) and mechanistic insight (*y*-axis). Black-box DL decoders (e.g., BCI, handwriting) achieve high predictive accuracy but provide limited explanatory power. Post hoc interpretability methods offer partial mechanistic clues, while hybrid DL architectures embedding biophysical priors may balance prediction with insight. In contrast, classical mechanistic models (e.g., optimal feedback control, dynamical systems, λ-model) offer high explanatory value but are often less generalizable in prediction. The figure highlights that predictive success and mechanistic understanding are not synonymous, and that their integration remains an open challenge in motor control research.

5. The Illusion of Understanding

While Section 2 characterized the intrinsic opacity of DL models—the architectural and algorithmic features that make them "black boxes"—the challenges do not end at the technical level. An equally serious problem arises from how the output of these models are interpreted within neuroscience. Beyond structural opacity, researchers and the broader community are vulnerable to perceptual and conceptual illusions that can distort scientific inference. These illusions include the confusion of correlation with causation, the anthropomorphization of artificial networks, and the seductive appeal of complexity. Whereas Section 2 emphasized why DL is inherently difficult to interpret, the present section focuses on the epistemic risks of misreading or overinterpreting DL outputs.

5.1. Confusing Correlation with Causation

One of the most significant risks of using black box models in neuroscience is the tendency to confuse statistical association with causal understanding. DL models are exceptionally good at finding correlations in high-dimensional data, but these correlations may not reflect causal relationships or mechanistic insights. When a DL model successfully predicts neural activity or behavior, it is tempting to conclude that it has captured something meaningful about the underlying biological processes. However, the model may have learned to exploit statistical regularities that are entirely incidental to the actual mechanisms of interest. This is particularly problematic in neuroscience, where the goal is typically to

understand causal relationships and mechanistic processes rather than merely to make predictions. A recent DL application in handwriting trajectory decoding illustrates this risk (Box 3).

Box 3. Handwriting/trajectory decoding.

Recent DL applications have achieved remarkable performance in brain-to-text communication by decoding handwriting trajectories from motor cortical activity. While methodologically groundbreaking, these models do not uncover a "neural code for handwriting." Instead, they provide powerful engineering solutions without revealing the causal computations of motor cortex. Overinterpretation of such outputs risks conflating predictive power with mechanistic insight.

5.2. The Anthropomorphization of Artificial Networks

Another subtle but important issue is the tendency to anthropomorphize AI, i.e., to attribute human-like or brain-like properties to systems that may operate through fundamentally different principles [68]. When researchers describe what a network "learns" or how it "represents" information, they often implicitly assume that these processes are analogous to biological learning and representation. This anthropomorphization can lead to misleading interpretations of model behavior. Features or representations that emerge in artificial networks may be purely computational artifacts with no biological analog. The language used to describe these features may suggest biological relevance where none exists [69]. It is crucial to recognize that AI has been modeled from biological systems, not the other way around.

5.3. The Seductive Appeal of Complexity

The complexity and sophistication of DL models can create an illusion of depth and insight. The fact that these models involve millions of parameters and complex nonlinear transformations [7,8] can make their outputs seem profound and meaningful, even when they may simply be sophisticated forms of curve fitting. This seductive appeal of complexity can lead researchers to overinterpret the results of DL models and to attribute more biological significance to their outputs than is warranted, particularly amidst a scientific environment that thrives on media hype [70,71]. The impressive performance of these models in prediction tasks can mask their fundamental limitations as tools for scientific understanding of high-complexity processing-driven phenomena.

6. Looking Forward

6.1. The Need for Epistemic Humility

The issues raised in this paper suggest the need for greater epistemic humility in the application of DL to neuroscience. Researchers should be more cautious about claims regarding what AI models reveal about brain function and more explicit about the limitations of these approaches. This humility should extend to the interpretation of model results, the claims made about biological relevance, and the implications drawn for understanding neural computation. It should also inform decisions about when and how to use AI models in neuroscience research.

6.2. Rethinking Evaluation Criteria

The neuroscience community may need to rethink how it evaluates computational models. Rather than focusing primarily on predictive performance, evaluation criteria should also consider interpretability, biological plausibility, and mechanistic insight. Models that perform slightly worse but offer greater understanding may be more valuable for scientific progress than opaque models with superior performance, at least for present motor control research. This shift in evaluation criteria would require changes in publica-

tion standards and the development of new metrics for assessing the scientific value of computational models beyond their predictive accuracy.

To move from exhortation to practice, we provide a concise, checklist-style rubric (Table 2) that authors, reviewers, and editors can apply when evaluating AI contributions in motor control. The rubric complements the philosophical criteria discussed in Sections 4 and 5 by translating them into operational items and minimal reporting standards.

Table 2. Evaluation rubric for AI models in motor control research.

Criterion	Question	Scoring (0–2)/Notes
Task type *(categorical)*	Is the study aimed at engineering-style prediction/decoding, mechanistic inference, or hybrid use?	Classification only (no score)
Mechanistic commitments	Are mechanistic claims explicit, tied to identifiable entities/activities and to a theoretical framework (e.g., optimal feedback control, internal models, dynamical systems, equilibrium-point, ecological)?	0 = None/implicit 1 = Partial/indirect 2 = Explicit & testable
Causal testability	Does the study include interventions, counterfactual predictions, or ablation tests that could falsify mechanistic claims?	0 = No 1 = Limited/indirect 2 = Direct & rigorous
Biophysical plausibility	Do model states/parameters map plausibly onto known neurophysiology/biomechanics (population dynamics, spinal circuits, local learning rules)?	0 = None 1 = Coarse analogy 2 = Substantive mapping
Generalization tests	Does the model generalize beyond training data (new tasks, effectors, datasets, labs)?	0 = In-distribution only 1 = Limited 2 = Multiple rigorous tests

Notes: Task type is categorical only (not scored); items 2–5 apply if mechanistic inference is claimed. Recommended reporting (non-scored): transparency/interpretability, baselines & ablations, reproducibility. Core score meaning: 0–3 = predictive only; 4–7 = partial mechanistic alignment; 8–10 = strong mechanistic alignment. This rubric complements domain-specific standards (see Sections 3.3–3.7, 4 and 5).

Concretely, this shift would involve adopting evaluation metrics that explicitly balance accuracy with interpretability and causal insight. Potential alternatives include hybrid frameworks that embed mechanistic priors into DL architectures—for example, spiking neural networks [63], energy-based models constrained by cortical dynamics, or architectures implementing local synaptic learning rules [64]. Other constructive approaches include the use of counterfactual testing and causal interventions, and the systematic comparison of DL models with established biophysical or dynamical systems models.

By explicitly prioritizing interpretability, causal validity, and biological plausibility alongside predictive performance, the field can move beyond critique and towards constructive methodological innovation. Such standards would help reorient DL from being primarily predictive tools toward explanatory models that advance neuroscientific understanding.

6.3. Education and Training

The integration of AI into neuroscience also highlights the need for improved education and training in computational methods. Researchers need to understand not just how to apply these techniques but also their limitations and appropriate use cases. This includes understanding the epistemological issues discussed in this review and the ability to critically evaluate the scientific value of different modeling approaches [72]. Training programs should emphasize the importance of matching computational methods to scientific questions and the need to consider interpretability and mechanistic insight alongside predictive performance.

6.4. Developing New Theoretical Frameworks

The challenges previously identified suggest the need for new theoretical frameworks that can better integrate computational modeling with scientific understanding. These approaches should explicitly address the relationship between artificial and biological computation and provide principled methods for extracting scientific insights from computational models. Such frameworks might draw on philosophy of science, cognitive science, and systems biology to develop new approaches to model interpretation and validation. They should also consider the unique challenges posed by the complexity and opacity of biological systems themselves.

6.5. Fostering Interdisciplinary Collaboration

Addressing the epistemological challenges of AI in neuroscience will require collaboration between neuroscientists, computer scientists, biologists, cognitive scientists, philosophers of science, and other disciplines. This collaboration should focus on developing new methodological approaches, theoretical frameworks, and evaluation criteria that can better serve the goals of scientific understanding. Such collaboration should also help ensure that the development of new computational methods is guided by the specific needs and constraints of neuroscience research rather than simply adapting techniques developed for other domains.

7. Conclusions

The widespread adoption of DL in neuroscience represents both an opportunity and a challenge. While these models offer impressive computational capabilities and have shown success in various applications, their black box nature creates fundamental tensions with the goals of scientific understanding that are central to neuroscience. The core argument of this review is that the opacity of DL models makes them unsuitable for the kind of mechanistic understanding that neuroscience requires. While these models can identify patterns and make predictions, they cannot provide the causal insights and mechanistic explanations that are necessary for advancing our understanding of brain function. This does not mean that DL has no place in neuroscience. These models may be valuable for certain applications, such as data preprocessing, feature extraction, or clinical prediction tasks where interpretability is less critical. However, the field must be more careful about distinguishing between applications where black box models are appropriate and those where mechanistic understanding is essential. The path forward requires a more nuanced approach to computational modeling in neuroscience—one that carefully matches methods to scientific questions, maintains appropriate epistemic humility about what different approaches can reveal, and continues to prioritize mechanistic understanding alongside predictive performance. As neuroscience continues to generate increasingly large and complex datasets, the temptation to rely on powerful but opaque computational methods will only grow. However, the field must resist the seductive appeal of black box approaches when they conflict with the fundamental goals of scientific understanding. Only by maintaining a clear focus on mechanistic insight and causal understanding can computational neuroscience fulfill its promise of advancing our knowledge of the brain. The epistemological challenges identified are not merely philosophical concerns but practical issues that affect the direction and progress of neuroscience research. By acknowledging these challenges and working to address them, the field can develop more effective approaches to computational modeling that truly serve the goals of scientific understanding rather than merely achieving impressive performance metrics. The future of computational neuroscience depends on finding ways to harness the power of modern computational methods while maintaining the interpretability and mechanistic insight that

are essential for scientific progress. This will require continued dialogue between different disciplines, ongoing development of new methodological approaches, and a commitment to the fundamental values of scientific inquiry that have driven progress in neuroscience for decades.

Author Contributions: Conceptualization, N.D.; writing—original draft preparation, N.D. and F.P.; writing—review and editing, N.D., F.P., L.P., S.S., M.F. and V.F.; visualization, M.F., V.F., L.P. and S.S.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
ANN	Artificial Neural Network

References

1. Dzobo, K.; Adotey, S.; Thomford, N.E.; Dzobo, W. Integrating Artificial and Human Intelligence: A Partnership for Responsible Innovation in Biomedical Engineering and Medicine. *Omics J. Integr. Biol.* **2020**, *24*, 247–263. [CrossRef] [PubMed]
2. Macpherson, T.; Churchland, A.; Sejnowski, T.; DiCarlo, J.; Kamitani, Y.; Takahashi, H.; Hikida, T. Natural and Artificial Intelligence: A Brief Introduction to the Interplay Between AI and Neuroscience Research. *Neural Netw.* **2021**, *144*, 603–613. [CrossRef] [PubMed]
3. Onciul, R.; Tataru, C.-I.; Dumitru, A.V.; Crivoi, C.; Serban, M.; Covache-Busuioc, R.-A.; Radoi, M.P.; Toader, C. Artificial Intelligence and Neuroscience: Transformative Synergies in Brain Research and Clinical Applications. *J. Clin. Med.* **2025**, *14*, 550. [CrossRef]
4. Tekin, U.; Dener, M. A Bibliometric Analysis of Studies on Artificial Intelligence in Neuroscience. *Front. Neurol.* **2025**, *16*, 1474484. [CrossRef]
5. Ali, S.; Abuhmed, T.; El-Sappagh, S.; Muhammad, K.; Alonso-Moral, J.M.; Confalonieri, R.; Guidotti, R.; Del Ser, J.; Díaz-Rodríguez, N.; Herrera, F. Explainable Artificial Intelligence (XAI): What We Know and What Is Left to Attain Trustworthy Artificial Intelligence. *Inf. Fusion* **2023**, *99*, 101805. [CrossRef]
6. Hassija, V.; Chamola, V.; Mahapatra, A.; Singal, A.; Goel, D.; Huang, K.; Scardapane, S.; Spinelli, I.; Mahmud, M.; Hussain, A. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cogn. Comput.* **2024**, *16*, 45–74. [CrossRef]
7. Bahri, Y.; Kadmon, J.; Pennington, J.; Schoenholz, S.S.; Sohl-Dickstein, J.; Ganguli, S. Statistical Mechanics of Deep Learning. *Annu. Rev. Condens. Matter Phys.* **2020**, *11*, 501–528. [CrossRef]
8. Nalisnick, E.; Smyth, P.; Tran, D. A Brief Tour of Deep Learning from a Statistical Perspective. *Annu. Rev. Stat. Appl.* **2023**, *10*, 219–246. [CrossRef]
9. Vyas, S.; Golub, M.D.; Sussillo, D.; Shenoy, K.V. Computation Through Neural Population Dynamics. *Annu. Rev. Neurosci.* **2020**, *43*, 249–275. [CrossRef]
10. Aliferis, C.; Simon, G. Lessons Learned from Historical Failures, Limitations and Successes of AI/ML in Healthcare and the Health Sciences. Enduring Problems, and the Role of Best Practices. In *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*; Simon, G.J., Aliferis, C., Eds.; Springer International Publishing: Cham, Switzerland, 2024; pp. 543–606, ISBN 978-3-031-39355-6.
11. Kanwisher, N.; Khosla, M.; Dobs, K. Using Artificial Neural Networks to Ask 'Why' Questions of Minds and Brains. *Trends Neurosci.* **2023**, *46*, 240–254. [CrossRef]

12. Ji, J.; Qiu, T.; Chen, B.; Zhang, B.; Lou, H.; Wang, K.; Duan, Y.; He, Z.; Vierling, L.; Hong, D.; et al. AI Alignment: A Comprehensive Survey. *arXiv* **2025**. [CrossRef]
13. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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*; ACM: San Francisco, CA, USA, 2016; pp. 1135–1144.
14. Asatiani, A.; Malo, P.; Nagbøl, P.R.; Penttinen, E.; Rinta-Kahila, T. Challenges of Explaining the Behavior of Black-Box AI Systems. *MIS Q. Exec.* **2020**, *19*, 259–278. [CrossRef]
15. Hao, X.; Zhang, G.; Ma, S. Deep Learning. *Int. J. Semant. Comput.* **2016**, *10*, 417–439. [CrossRef]
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
17. Janiesch, C.; Zschech, P.; Heinrich, K. Machine Learning and Deep Learning. *Electron. Mark.* **2021**, *31*, 685–695. [CrossRef]
18. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**. [CrossRef]
19. Rai, A. Explainable AI: From Black Box to Glass Box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141. [CrossRef]
20. Longo, L.; Brcic, M.; Cabitza, F.; Choi, J.; Confalonieri, R.; Ser, J.D.; Guidotti, R.; Hayashi, Y.; Herrera, F.; Holzinger, A.; et al. Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions. *Inf. Fusion* **2024**, *106*, 102301. [CrossRef]
21. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *arXiv* **2018**. [CrossRef]
22. Barrett, D.G.; Morcos, A.S.; Macke, J.H. Analyzing Biological and Artificial Neural Networks: Challenges with Opportunities for Synergy? *Curr. Opin. Neurobiol.* **2019**, *55*, 55–64. [CrossRef]
23. Abusitta, A.; Li, M.Q.; Fung, B.C.M. Survey on Explainable AI: Techniques, Challenges and Open Issues. *Expert Syst. Appl.* **2024**, *255*, 124710. [CrossRef]
24. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
25. Craver, C.F.; Darden, L. Mechanisms in Biology. Introduction. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 233–244. [CrossRef]
26. Bechtel, W.; Abrahamsen, A. Explanation: A Mechanist Alternative. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **2005**, *36*, 421–441. [CrossRef] [PubMed]
27. Casadevall, A.; Fang, F.C. Mechanistic Science. *Infect. Immun.* **2009**, *77*, 3517–3519. [CrossRef]
28. Nielsen, I.E.; Dera, D.; Rasool, G.; Ramachandran, R.P.; Bouaynaya, N.C. Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks. *IEEE Signal Process. Mag.* **2022**, *39*, 73–84. [CrossRef]
29. Brauwers, G.; Frasincar, F. A General Survey on Attention Mechanisms in Deep Learning. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 3279–3298. [CrossRef]
30. El Houda Dehimi, N.; Tolba, Z. Attention Mechanisms in Deep Learning: Towards Explainable Artificial Intelligence. In Proceedings of the 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS), El Oued, Algeria, 24–25 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–7.
31. Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K.R. Layer-Wise Relevance Propagation: An Overview. In *Lecture Notes in Computer Science*; Springer International Publishing: Cham, Switzerland, 2019; pp. 193–209, ISBN 978-3-030-28953-9.
32. Cuadra, C.; Wojnicz, W.; Kozinc, Z.; Latash, M.L. Perceptual and Motor Effects of Muscle Co-Activation in a Force Production Task. *Neuroscience* **2020**, *437*, 34–44. [CrossRef]
33. Latash, M.L.; Singh, T. *Neurophysiological Basis of Motor Control*, 3rd ed.; Human Kinetics: Champaign, IL, USA, 2024; ISBN 978-1-7182-0953-4.
34. Tanaka, H.; Miyakoshi, M.; Makeig, S. Dynamics of Directional Tuning and Reference Frames in Humans: A High-Density EEG Study. *Sci. Rep.* **2018**, *8*, 8205. [CrossRef]
35. Churchland, M.M.; Cunningham, J.P.; Kaufman, M.T.; Foster, J.D.; Nuyujukian, P.; Ryu, S.I.; Shenoy, K.V. Neural Population Dynamics During Reaching. *Nature* **2012**, *487*, 51–56. [CrossRef]
36. Latash, M.L.; Zatsiorsky, V.M.; Latash, M.L. *Biomechanics and Motor Control: Defining Central Concepts*; Elsevier/AP: London, UK; San Diego, CA, USA, 2016; ISBN 978-0-12-800384-8.
37. Feldman, A.G. Origin and Advances of the Equilibrium-Point Hypothesis. In *Progress in Motor Control: A Multidisciplinary Perspective*; Sternad, D., Ed.; Springer: Boston, MA, USA, 2009; pp. 637–643, ISBN 978-0-387-77064-2.
38. Gallego, J.A.; Perich, M.G.; Miller, L.E.; Solla, S.A. Neural Manifolds for the Control of Movement. *Neuron* **2017**, *94*, 978–984. [CrossRef]
39. Merton, P.A. Speculations on the Servo-Control of Movement. In *Ciba Foundation Symposium—The Spinal Cord*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 1953; pp. 247–260, ISBN 978-0-470-71882-7.
40. Seidler, R.D.; Noll, D.C.; Thiers, G. Feedforward and Feedback Processes in Motor Control. *NeuroImage* **2004**, *22*, 1775–1783. [CrossRef]

41. Malone, L.A.; Hill, N.M.; Tripp, H.; Zipunnikov, V.; Wolpert, D.M.; Bastian, A.J. The Control of Movement Gradually Transitions from Feedback Control to Feedforward Adaptation Throughout Childhood. *npj Sci. Learn.* **2025**, *10*, 13. [CrossRef]
42. Favela, L.H.H. *The Ecological Brain*; Routledge: New York, NY, USA, 2024; ISBN 978-0-367-44471-6.
43. Blau, J.J.C.; Wagman, J.B. *Introduction to Ecological Psychology: A Lawful Approach to Perceiving, Acting, and Cognizing*, 1st ed.; Routledge: New York, NY, USA, 2022; ISBN 978-1-003-14569-1.
44. Gibson, J.J. *The Ecological Approach to Visual Perception*; Psychology Press: New York, NY, USA, 1979.
45. Alexander, G.E.; Crutcher, M.D. Neural Representations of the Target (Goal) of Visually Guided Arm Movements in Three Motor Areas of the Monkey. *J. Neurophysiol.* **1990**, *64*, 164–178. [CrossRef]
46. Alexander, G.E.; Crutcher, M.D. Preparation for Movement: Neural Representations of Intended Direction in Three Motor Areas of the Monkey. *J. Neurophysiol.* **1990**, *64*, 133–150. [CrossRef]
47. Georgopoulos, A.; Kalaska, J.; Caminiti, R.; Massey, J. On the Relations Between the Direction of Two-Dimensional Arm Movements and Cell Discharge in Primate Motor Cortex. *J. Neurosci.* **1982**, *2*, 1527–1537. [CrossRef]
48. Michaels, J.A.; Dann, B.; Scherberger, H. Neural Population Dynamics During Reaching Are Better Explained by a Dynamical System than Representational Tuning. *PLoS Comput. Biol.* **2016**, *12*, e1005175. [CrossRef]
49. Willett, F.R.; Avansino, D.T.; Hochberg, L.R.; Henderson, J.M.; Shenoy, K.V. High-Performance Brain-to-Text Communication via Handwriting. *Nature* **2021**, *593*, 249–254. [CrossRef]
50. Yamins, D.L.K.; DiCarlo, J.J. Using Goal-Driven Deep Learning Models to Understand Sensory Cortex. *Nat. Neurosci.* **2016**, *19*, 356–365. [CrossRef]
51. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu. Rev. Vis. Sci.* **2015**, *1*, 417–446. [CrossRef]
52. Sussillo, D.; Churchland, M.M.; Kaufman, M.T.; Shenoy, K.V. A Neural Network That Finds a Naturalistic Solution for the Production of Muscle Activity. *Nat. Neurosci.* **2015**, *18*, 1025–1033. [CrossRef]
53. Pandarinath, C.; O'Shea, D.J.; Collins, J.; Jozefowicz, R.; Stavisky, S.D.; Kao, J.C.; Trautmann, E.M.; Kaufman, M.T.; Ryu, S.I.; Hochberg, L.R.; et al. Inferring Single-Trial Neural Population Dynamics Using Sequential Auto-Encoders. *Nat. Methods* **2018**, *15*, 805–815. [CrossRef] [PubMed]
54. Marblestone, A.H.; Wayne, G.; Kording, K.P. Toward an Integration of Deep Learning and Neuroscience. *Front. Comput. Neurosci.* **2016**, *10*, 94. [CrossRef] [PubMed]
55. Woodward, J. *Making Things Happen: A Theory of Causal Explanation*; Oxford University Press: Oxford, UK, 2004; ISBN 978-0-19-515527-3.
56. Lundy-Ekman, L. *Neuroscience: Fundamentals for Rehabilitation*, 4th ed.; Elsevier Inc.: Amsterdam, The Netherlands, 2013; ISBN 978-1-4557-0643-3.
57. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. *Nature* **2017**, *542*, 115–118. [CrossRef]
58. Heo, J.; Yoon, J.G.; Park, H.; Kim, Y.D.; Nam, H.S.; Heo, J.H. Machine Learning–Based Model for Prediction of Outcomes in Acute Stroke. *Stroke* **2019**, *50*, 1263–1265. [CrossRef]
59. Scott, S.H. Optimal Feedback Control and the Neural Basis of Volitional Motor Control. *Nat. Rev. Neurosci.* **2004**, *5*, 532–545. [CrossRef]
60. Todorov, E.; Jordan, M.I. Optimal Feedback Control as a Theory of Motor Coordination. *Nat. Neurosci.* **2002**, *5*, 1226–1235. [CrossRef]
61. Kawato, M. Internal Models for Motor Control and Trajectory Planning. *Curr. Opin. Neurobiol.* **1999**, *9*, 718–727. [CrossRef]
62. De Regt, H.W.; Dieks, D. A Contextual Approach to Scientific Understanding. *Synthese* **2005**, *144*, 137–170. [CrossRef]
63. Tavanaei, A.; Ghodrati, M.; Kheradpisheh, S.R.; Masquelier, T.; Maida, A. Deep Learning in Spiking Neural Networks. *Neural Netw.* **2019**, *111*, 47–63. [CrossRef] [PubMed]
64. Konishi, M.; Igarashi, K.M.; Miura, K. Biologically Plausible Local Synaptic Learning Rules Robustly Implement Deep Supervised Learning. *Front. Neurosci.* **2023**, *17*, 1160899. [CrossRef]
65. Aliferis, C.; Simon, G. Overfitting, Underfitting and General Model Overconfidence and Under-Performance Pitfalls and Best Practices in Machine Learning and AI. In *Artificial Intelligence and Machine Learning in Health Care and Medical Sciences: Best Practices and Pitfalls*; Simon, G.J., Aliferis, C., Eds.; Springer International Publishing: Cham, Switzerland, 2024; pp. 477–524, ISBN 978-3-031-39355-6.
66. Shao, F.; Shen, Z. How Can Artificial Neural Networks Approximate the Brain? *Front. Psychol.* **2023**, *13*, 970214. [CrossRef] [PubMed]
67. Sinz, F.H.; Pitkow, X.; Reimer, J.; Bethge, M.; Tolias, A.S. Engineering a Less Artificial Intelligence. *Neuron* **2019**, *103*, 967–979. [CrossRef]
68. Placani, A. Anthropomorphism in AI: Hype and Fallacy. *AI Ethics* **2024**, *4*, 691–698. [CrossRef]
69. Barrow, N. Anthropomorphism and AI Hype. *AI Ethics* **2024**, *4*, 707–711. [CrossRef]

70. Abdollahi, M.; Gasparyan, A.Y.; Saeidnia, S. The Urge to Publish More and Its Consequences. *DARU J. Pharm. Sci.* **2014**, *22*, 53. [CrossRef]
71. Singhal, S.; Kalra, B.S. Publication Ethics: Role and Responsibility of Authors. *Indian J. Gastroenterol.* **2021**, *40*, 65–71. [CrossRef]
72. Wang, X.-J.; Hu, H.; Huang, C.; Kennedy, H.; Li, C.T.; Logothetis, N.; Lu, Z.-L.; Luo, Q.; Poo, M.; Tsao, D.; et al. Computational Neuroscience: A Frontier of the 21st Century. *Natl. Sci. Rev.* **2020**, *7*, 1418–1422. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.