A Mechanistic Explanatory Strategy for XAI

Forthcoming in Müller, V. C., Dewey, A. R., Dung, L., & Löhr, G. (Eds.), *Philosophy of Artificial Intelligence: The State of the Art*, Synthese Library, Berlin: Springer Nature. Please cite the published version.

Marcin Rabiza [1,2]

[1] Institute for Philosophy, Leiden University, Leiden, The Netherlands

[2] Institute of Philosophy and Sociology, Polish Academy of Sciences, Warsaw, Poland

marcin.rabiza@gssr.edu.pl

https://orcid.org/0000-0001-6217-6149

Abstract: Despite significant advancements in XAI, scholars note a persistent lack of solid conceptual foundations and integration with broader scientific discourse on explanation. In response, emerging XAI research draws on explanatory strategies from various sciences and philosophy of science literature to fill these gaps. This paper outlines a mechanistic strategy for explaining the functional organization of deep learning systems, situating recent advancements in AI explainability within a broader philosophical context. According to the mechanistic approach, the explanation of opaque AI systems involves identifying mechanisms that drive decision-making. For deep neural networks, this means discerning functionally relevant components—such as neurons, layers, circuits, or activation patterns—and understanding their roles through decomposition, localization, and recomposition. Proof-of-principle case studies from image recognition and language modeling align these theoretical approaches with the latest research from AI labs like OpenAI and Anthropic. This research suggests that a systematic approach to studying model organization can reveal elements that simpler (or "more modest") explainability techniques might miss, fostering more thoroughly explainable AI. The paper concludes with a discussion on the epistemic relevance of the mechanistic approach positioned in the context of selected philosophical debates on XAI.

1 Introduction

In recent years, *deep learning* has emerged as the dominant approach in *artificial intelligence* (AI). Its algorithms are characterized by "black box" functions that are too complex to fully understand, making it difficult to determine how specific computations based on certain inputs lead to particular predictions. To restore trust in automated decision-making, researchers focus on the *explainable AI* (XAI) research program, which aims to achieve transparency, interpretability, and explainability to validate the decision-making processes of opaque AI systems, making model behavior understandable to stakeholders with various epistemic needs.[1]

[1] *Interpretability* is often defined as "the degree to which an observer can understand the cause of a decision" (Miller, 2019, p. 8). Explanation, therefore, is one mode through which an observer may obtain such understanding. In the machine learning literature, the terms "interpretability"

Despite significant technical advancements in XAI (e.g., Montavon et al., 2018; Linardatos et al., 2021), scholars highlight various shortcomings in its conceptual foundations and a lack of integration with the broader scientific discourse on explanation. While many current XAI methods excel in producing localized explanations, they often fall short of providing a comprehensive understanding of the complex mechanisms governing AI systems, which is crucial in high-stakes decision-making contexts (Mittelstadt et al., 2019). Furthermore, the dominant technology-centered approach in AI explanations has largely ignored substantial theoretical and empirical contributions from fields like philosophy, psychology, and cognitive sciences, leaving a significant research area underexplored (Miller, 2019). In response, an emerging strand of fundamental XAI research is drawing on coordinated explanatory strategies from various scientific disciplines and the philosophy of science literature to address these conceptual gaps (e.g., Miller et al., 2017; Lipton, 2018; Mittelstadt et al., 2019; Páez, 2019; Zednik, 2019; Zerilli et al., 2019; Erasmus et al., 2021; Watson & Floridi, 2020; Beisbart & Räz, 2022).

This paper outlines a mechanistic strategy for explaining the functional organization of deep learning systems, situating recent advancements in XAI methods—especially the mechanistic interpretability movement—within a broader philosophical context. To this end, I draw upon the tradition of the new mechanism in the philosophy of science while leveraging examples derived from XAI engineering practices.

The structure of the paper is as follows. Section 2 introduces the concept of a mechanism based on a minimal neomechanistic theory of explanation and argues for its applicability to XAI. Subsequently, Section 3 presents a mechanistic interpretation of deep learning, explaining the operations of deep neural networks by identifying decision-making mechanisms through discovery heuristics of decomposition, localization, and recomposition. Section 4 then utilizes proof-of-principle case studies from image recognition and language modeling to align these theoretical approaches with the latest research from leading AI labs like OpenAI and Anthropic. Section 5 concludes with a brief discussion of the epistemic relevance of the mechanistic approach in the context of XAI.

---

and "explainability" are often used interchangeably (e.g., Molnar, 2022, §3)—a convention I will follow for now until I introduce a specific understanding of *interpretable* and *comprehensible* systems from Doran et al. (2017) later in the paper. However, it should be noted that in some contexts, researchers differentiate these two notions (e.g., Dorsch & Moll, Forthcoming).

2 Neomechanistic Theory of Explanation

In scientific discourse, one prominent approach is characterized by the neomechanistic theory of explanation, which emphasizes the logic of "explaining *why* by explaining *how*" (Bechtel & Abrahamsen, 2005, p. 422). According to the new mechanists, explaining why something happens often involves identifying the underlying mechanisms that give rise to observed behavior. *Mechanisms* are identified by the phenomena they produce (Illari & Williamson, 2012), their functional roles (Machamer, Darden, & Craver, 2000; hereinafter MDC), and by their operating "parts" and "interactions." (Bechtel & Abrahamsen, 2005). For example, Glennan (1996, p. 52) defines a mechanism as "a complex system which produces that behavior by the interaction of a number of parts according to direct causal laws." According to MDC (2000, p. 3), mechanisms are identified and individuated by the "activities" and "entities" that constitute them, as well as their functional roles, and setup and termination conditions.

Entities are components or parts defined by their properties—such as location, structure, and orientation— that engage in activities based on specific characteristics. *Activities* are temporal producers of change, characterized by aspects such as spatiotemporal location, rate, duration, types of entities involved, and other operational properties. The roles that entities play through their activities are considered their *functions* within the mechanism. The specific organization of these elements determines how collectively they produce the observed phenomenon (MDC, 2000).

Mechanistic explanation begins by characterizing the phenomenon under study and then identifying the mechanisms responsible for it. According to Bechtel and Abrahamsen (2005), to explain a mechanism, scientists must pinpoint its components, understand the functions these parts perform, and determine their organization to produce the phenomenon. This process relies on scientific discovery methods, incorporating heuristic strategies such as decomposition and localization (Bechtel & Richardson, 2010).

Decomposition involves breaking the mechanism into its structural or functional components. Structural decomposition focuses on the physical aspects of parts like size or shape, while functional decomposition looks at the parts' roles, causal powers, and overall contributions to the mechanism (Piccinini & Craver, 2011). Functional decomposition assumes that the system's behavior results from its sub-functions. Structural decomposition further breaks down these functions into their physical components. This process starts with hypothetical component parts, refining the breakdown as more is learned about the system's

operations, with both types of decomposition eventually integrating to form a complete explanation.

Localization complements the decomposition process by mapping component operations onto their respective parts. While decomposition involves breaking down the mechanism into parts and operations, localization identifies activities from the task decomposition and links them to component behaviors or capabilities (Wright & Bechtel, 2007). Sometimes, physical components can be directly identified, but often their existence is inferred using functional tools without direct observation. Bechtel and Abrahamsen (2005) note that even inferred components are treated as essential parts of the mechanism. Localization involves a genuine commitment to the functions identified in the task decomposition and the use of appropriate methods to demonstrate that something within the system is performing each of these functions.

Mechanistic explanations describe the relevant entities, their properties, and the activities that connect them, by demonstrating how actions at one stage influence those at successive stages. Glennan (2017) notes that mechanistic models can be depicted through diagrams accompanied by linguistic explanations. These diagrams typically illustrate spatial relations and structural features of the mechanism, with related activities depicted as labeled arrows (see Figure 2.1). Although the basic arrangement of a mechanism might be linear, it can also include more complex structures like forks, joins, cycles, and non-linear interactions.

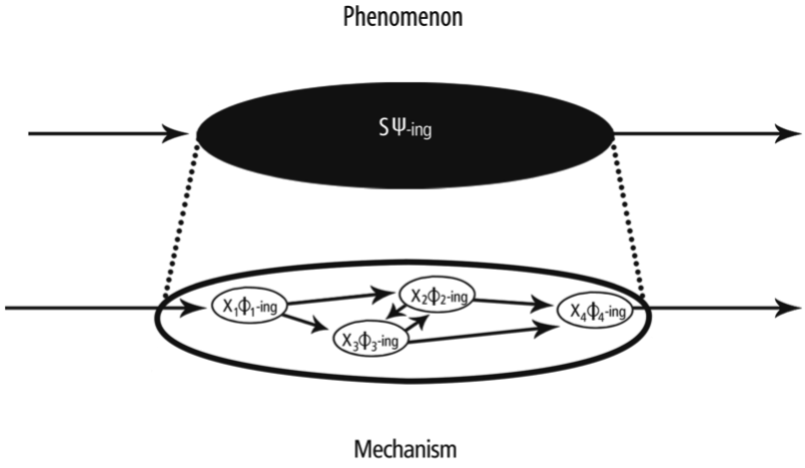

Fig. 2.1 A diagrammatic representation of a mechanism (reproduced from Craver, 2007). At the top is the phenomenon: some system *S* engaged in behavior ψ. Beneath it are the parts (the *X*s) and their activities (the φs) organized together.

As Figure 2.1 might suggest, mechanisms form nested, multilevel hierarchies, in which lower-level entities and activities serve as components for higher-level phenomena, effectively becoming mechanisms themselves. These hierarchies have a finite structure and do not decompose indefinitely; the lowest level of description is determined by practical

considerations and stakeholder interests. Although all mechanisms are ultimately based on fundamental, non-causal physical laws, seeking explanations at this level is typically neither practical nor beneficial. Instead, explanations tend to bottom out at components that are fundamental or unproblematic within a specific scientific discipline or explanatory practice (MDC, 2000).

In the optional stage of mechanism discovery, as described by Bechtel and Abrahamsen (2013), scientists may *recompose* what they have learned about the functional parts into an explanatory model, such as a mathematical or computational model. The goal is to create a model from a catalog of entities likely to be causally relevant to a phenomenon based on their identified internal division of labor.

The mechanistic explanatory strategy that emerges from this description resembles something like a *causal narrative*, in the sense that it outlines sequences of events involving entities interacting with each other, illustrating how their spatial and temporal arrangement produces or sustains the explanandum (Glennan, 2017). An important question at hand is: How might this framework be adapted to analyze and explain opaque AI systems?

3 Mechanistic Interpretation of Deep Learning

3.1 Mechanistic Structure of Deep Neural Networks

The mechanistic explanatory strategy has been adopted across various scientific fields, being especially prevalent in neuroscience (Kostić & Halffman, 2023). Given the biological inspiration of *deep neural networks* (DNNs), along with their computational parallels and predictive capabilities akin to the brain's categorization processes (Schyns et al., 2022), it may be promising to apply mechanistic principles to AI systems using opaque deep learning algorithms. Indeed, Kästner and Crook (2024) argue that as AI systems grow in complexity, they should be analyzed similarly to biological organisms, emphasizing their functional organization. To this end, applying the mechanistic approach to DNNs could help us discover *how* these models work internally, illuminating not only how specific computations—given specific inputs—produce unique predictive patterns, but perhaps also explaining the models themselves in a holistic way.

In the neomechanistic framework, explaining AI systems entails identifying the mechanisms behind their decision-making processes using discovery heuristics of decomposition, localization, and recomposition. The goal is to understand the properties driving

behavior and how these are orchestrated through component interactions (see Figure 3.1). By dissecting neural networks into comprehensible components, researchers can better grasp each part's function and structure, gaining insights into the network's overall behavior.

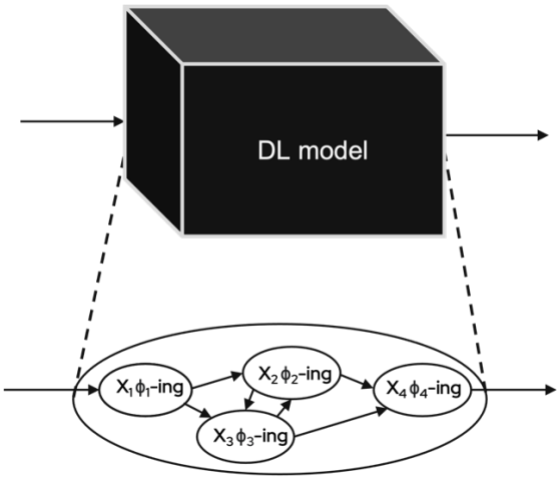

Fig. 3.1 Schematic representation: internal structure of the mechanism of the deep learning model analyzed in the neomechanistic framework.

The first step in this endeavor involves identifying the correct components for analysis—the candidates for the AI system's *epistemically relevant elements* (EREs) (Humphreys, 2009; Kästner & Crook, 2024). Neurons, while fundamental computational units of neural networks, often fall short as effective units for human interpretation. Despite performing straightforward arithmetic, individual neuron-like units viewed in isolation fail to clearly demonstrate how their functions contribute to the network's overall behavior. Consequently, researchers seek other network components as potential EREs that could offer more comprehensible units of analysis—robust patterns that sustain system behavior and are pertinent to explanatory functions (Kästner & Crook, 2024).

In principle, this approach should be compatible with deep learning, as DNNs are built from causally or functionally relevant components. These include entities (medium-independent vehicles) and activities (the manipulations these vehicles undergo) which are central to the computational mechanisms of deep learning. While various types of networks exist, they share common entities such as neurons, connections, filters, circuits, or features.[2] DNNs develop their functional organization through automated training processes. Although models are processed as multidimensional arrays of numbers with mathematical operations

[2] It might not be clear in what way network structures qualify as mechanistic entities. While this issue deserves a more detailed discussion, for the purpose of this paper, I will assume that neurons are entities in the sense that they are mathematical abstractions of certain components ultimately belonging to the underlying physical hardware. Similar reasoning applies to activities, which eventually bottom out at the level of physical phenomena.

defined over vectors in programming languages such as Python, computer scientists interested in *mechanistic interpretability*—a particular approach to XAI focused on deciphering the internal workings of machine learning models—recognize that "neural network parameters are in some sense a binary computer program which runs on one of the exotic virtual machines we call a neural network architecture" (Olah, 2022). DNN components can be identified within such environments. During training, these entities engage in activities such as activation, back-propagation, and error minimization, triggered by properties like incoming signals exceeding certain thresholds. This interaction fosters the development of specialized roles within the system, often unforeseen by programmers, contributing to the emergence of observed behavior and supporting the assumptions of mechanistic interpretability.

An example of this "mechanistic compatibility" can be seen in deep *convolutional neural networks* (CNNs), primarily used in image recognition and computer vision, which somewhat mimic the organization of the animal visual cortex. During training, a CNN processes a two-dimensional labeled image to generate weights that encode extracted data patterns. CNNs employ organized entities—such as neurons, larger neuronal circuits, convolutional kernels (filters), or various kinds of layers—along with activities like convolutions, ReLU and softmax activations, and pooling, orchestrating complex mechanisms of feature extraction and image classification (see Figure 3.2).

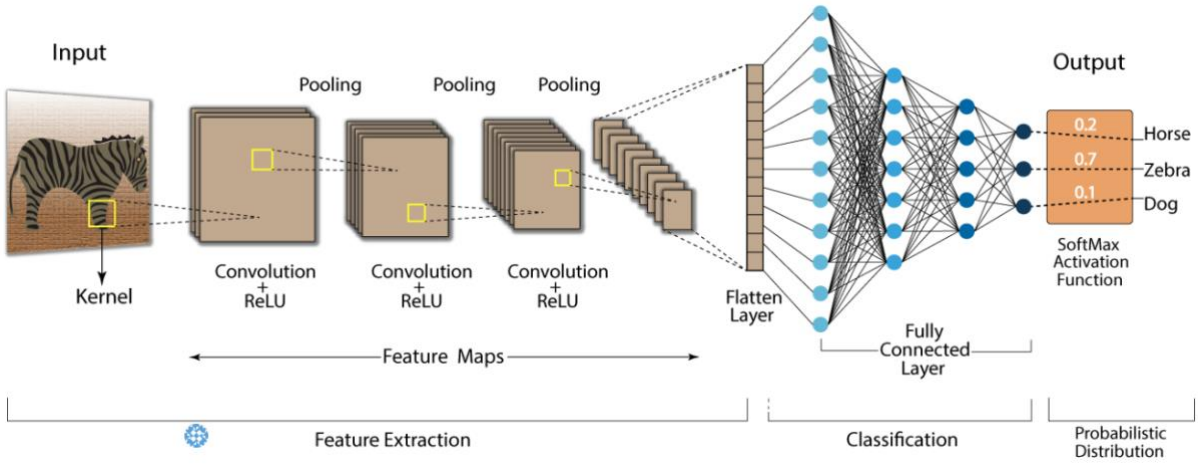

Fig. 3.2 Architecture of a deep convolutional neural network (reproduced from Shahriar, 2023).

The inference phase of image recognition begins with *setup conditions* that include initializing model parameters. An input image is then introduced and subjected to a series of transformations across multiple layers to extract and refine features. *Intermediate activities* of this process involve convolution, activation, and pooling. Convolutional kernels slide across the input signal, detecting features similar to biological neural networks' receptive fields and

generating feature maps for subsequent layers. ReLU activations eliminate negative values, activating only nodes that exceed a certain threshold. Pooling layers reduce data dimensionality by merging outputs from neuron clusters into single neurons, using hierarchical patterns to evolve simpler features into more complex ones. As processing continues, layers represent diverse image features such as edges, lines, and curves, with higher layers capturing more complex, "meaningful" and abstract shapes. The process culminates in fully connected layers in which the softmax function transforms raw outputs into class probability scores, marking the *termination condition* of the inference phase. This sequence, maintained under stable and consistent conditions, demonstrates the deterministic regularity characteristic of genuine mechanisms.

In this setup, CNN mechanisms form multilevel hierarchies, in which lower-level entities and activities act as enabling components for higher-level phenomena, thus illustrating the mechanistic nature of their internal organization. For example, input convolution is crucial for feature mapping, which, when applied iteratively, integrates into the overarching mechanism of image classification in fully connected layers.

3.2 Implementation of Mechanism Discovery Heuristics

Having explored the mechanistic structure of DNNs, it's important to consider how we can systematically apply discovery heuristics of decomposition and localization to dissect and understand the roles of EREs within these networks. In XAI practice, these strategies can be implemented through established analytical techniques tailored to specific applications. In computer vision, for example, input heatmapping and feature visualization techniques can be used to generate saliency maps that highlight specific pixels or regions in an input image that are highly predictive of the output. Such visualizations can also clarify the operations performed by DNN components across layers, thus aiding the functional decomposition of the network. Concrete examples include activation maximization, regularized optimization, network inversion, deconvolutional neural networks, network dissection-based visualization, or layer-wise relevance propagation (Yosinski et al., 2015; Qin et al., 2018; Montavon et al., 2018). These techniques allow researchers to observe how each network level transforms input into increasingly abstract and meaningful representations, which is crucial for dissecting complex mechanisms like face recognition into simpler components that recognize individual features such as ears, eyes, or noses. An example of hierarchical feature representations processed within a CNN is shown in Figure 3.3.

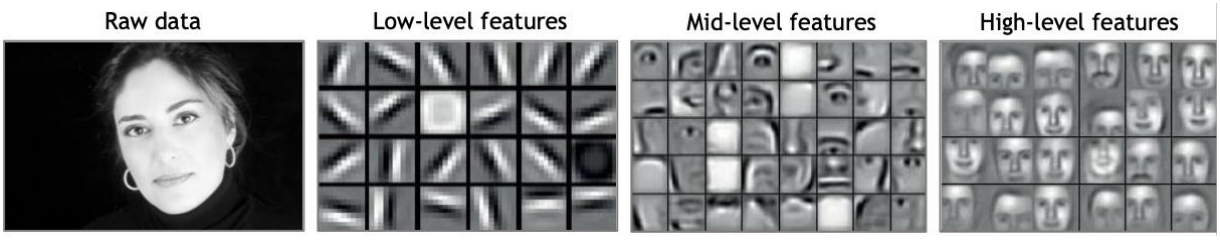

Fig. 3.3 Visualization of features on various layers of a CNN for input images of faces (reproduced from Karmakharm, 2018).

While saliency methods may not provide complete mechanistic explanations by themselves, they are instrumental in the mechanistic decomposition of DNNs by identifying distributed structures beyond individual neurons as potential EREs of AI decision-making mechanisms. Recent advancements by leading AI organizations such as OpenAI, Anthropic, Google, Redwood Research, ARC, and Conjecture demonstrate a similar interest in such "mechanistic" understanding, particularly within the AI safety community (e.g., Olah et al., 2017; Cammarata et al., 2021; Elhage et al., 2021; Chan et al., 2022; Christiano, 2022; Olsson et al., 2022; Bricken et al., 2023a; Cunningham et al., 2023; Conmy et al., 2023; Schwettmann et al., 2023). This trend is evident in the mechanistic interpretability movement, which aims to go beyond simple input-output analysis and examine the internal workings of AI models to enhance epistemic trust, aid in debugging, remove biases, and prevent models from "going rogue."

This approach is exemplified by the work of Cammarata et al. (2021) at OpenAI, who aim to reverse-engineer image classification neuron families known as *curve detectors* into understandable explanations and then implement the inferred algorithms into a new DNN from scratch. Their research particularly focuses on analyzing a curve detector circuit within the fifth convolutional layer of the InceptionV1 neural network. These circuits, which are sub-graphs within the network, are not predefined as distinct parts of the DNN's architecture but are integrated into the model's learned structure, emerging as functional units that neurons self-organize into during training.

To understand the functional organization of these circuits (termed the "mechanics of curve detectors" by the authors), the team employs decomposition and localization strategies using *decomposed feature visualization* to create a grid that illustrates amplified weights from an upstream layer to a downstream neuron of interest. By iteratively applying this technique to each neuron, labeling them, and grouping them, they categorize the neurons in the first five convolutional layers of InceptionV1 into layer-wise "families" that form the curve detection

mechanisms. Upon identifying curve detectors, the researchers traced their connections to discern how upstream neurons affected their activities by visualizing connection weights. Extending this visualization back to the input layer provided a detailed view of the interactions within the circuit, enabling them to classify the circuit as a mechanistic ERE distinct from the surrounding network (cf. Kästner & Crook, 2024, p. 10). The team then developed a high-level "circuit schematic" that details the primary components of curve detection and their functional organization, forming a clear narrative: "Gabor filters evolve into proto-lines, which assemble into lines and early curves, ultimately forming curves" (Cammarata et al., 2021).

Leveraging insights from the discovery process, the researchers recomposed a curve detection mechanism of InceptionV1 by manually configuring the weights of a blank DNN to replicate the identified neuron families and circuit interactions. They compared the behavior of the manually designed network with InceptionV1 using identical synthetic stimuli and analyzed responses with feature visualization and other XAI techniques. The experiments showed that the artificially recomposed curve detectors closely resembled those trained naturally. This evidence suggests that the functional decomposition of InceptionV1 was indeed successful and accurately reflects the mechanistic organization of curve detection circuits.

Anthropic, known for developing Claude—a large language model that rivals OpenAI's ChatGPT—has employed decomposition and localization strategies to break down language models into interpretable, structurally distinct functional components. Bricken et al. (2023a; 2023b), recognizing that individual neurons are not the most effective units for analysis, partly due to their polysemanticity, employ a *sparse autoencoder*—a type of weak dictionary learning algorithm—to identify better candidates for EREs in small transformer models. These units, termed "features," represent distinct patterns within the model and are essentially linear combinations of neuron activations.

In their study of a transformer language model, the researchers decomposed a layer with 512 neurons into over 4,000 features by training sparse autoencoders on multilayer perceptron activations from 8 billion data points. Each feature captured a unique concept, such as DNA sequences, legal language, HTTP requests, Hebrew text, and nutrition statements. To evaluate the interpretability of these features—that is, the degree to which humans can understand them—they conducted an assessment with a blinded human evaluator (Figure 3.4), validating the practical utility and clarity of the decomposed features.

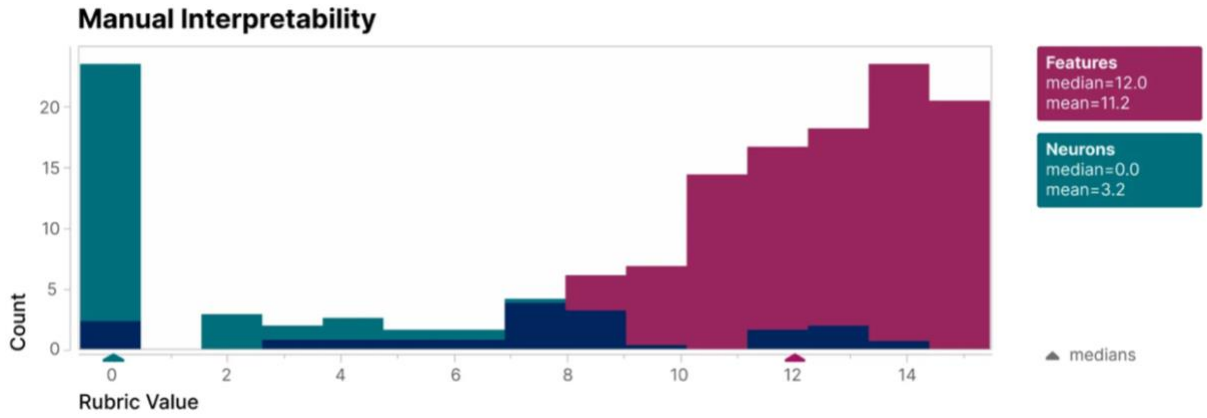

Fig. 3.4 Interpretability scores of the identified features (red) compared to the neurons (teal) (reproduced from Bricken et al., 2023b).

The results showed that the features were significantly more interpretable than individual neurons, revealing functional properties that were not apparent when examining neurons alone. Moreover, the authors conducted "autointerpretability" tests, in which a language model generated concise descriptions of the small model's features. The evaluation of these descriptions was based on another model's ability to predict a feature's activations from its description. The features consistently received higher scores than the neurons, confirming their coherent and stable interpretation and their impact on model behavior.

Decomposing the language model into features offers a targeted method for guiding models, in which the activation of specific features leads to predictable changes in behavior. Researchers also developed a "knob" to adjust the resolution at which the model is viewed and experimented with the number of features learned. They found that decomposing the model into a small set of features offers a coarse but clear view, while a larger set reveals more refined and subtle properties of the model. Additionally, these learned features have proven to be universal across a range of models, demonstrating the enhanced generalizability of this explanatory approach.

Overall, evidence from case studies supports the idea that a coordinated, systematic research agenda focused on uncovering the mechanistic organization of DNNs can provide explanations of the way that systems operate at various levels of their structure. The pursuit of mechanistic explanations through functional decomposition can reveal previously unknown EREs in opaque AI systems, which might remain obscured with non-coordinated individual explainability methods, ultimately leading to more thoroughly explainable AI.

4 Are Deep Neural Networks Genuine Mechanisms?

There may be several objections from orthodox mechanists regarding the characterization of DNNs as mechanisms, primarily questioning whether DNNs meet the criteria for genuine mechanisms as defined in neomechanistic literature. Despite possible skepticism, I argue that the discovery strategies proposed by neomechanistic philosophy—decomposition, localization, and recomposition—can help researchers at least partially open the deep learning black box.

A key standard in the general account of mechanistic explanation is the demand for completeness as causal models, necessitating that all causally relevant parts and operations be explicitly detailed without gaps or placeholders (Craver, 2007). A fully adequate mechanistic explanation must provide structural details at all levels of the mechanism, including components and activities that contribute to concrete computations. This requirement conflicts with the abstractness and medium-independence typical for computational explanations (Haimovici, 2013; Coelho Mollo, 2018). Under such scrutiny, AI models like DNNs might not qualify as genuine mechanisms because they typically represent abstract, formal specifications of computation that lack detailed structural information. DNNs are usually simulated through matrix operations rather than being implemented on physical nodes (except in rare cases involving one-to-one mapping, such as on neuromorphic hardware, cf. e.g., Schuman et al., 2022). To be fully mechanistically adequate, they would require additional specifications concerning physically instantiated computers: *instantiation blueprints* (Miłkowski, 2014).

On the other hand, there are some compelling reasons to treat DNNs as if they were mechanisms. While this paper cannot fully explore the debate over the ontic status of AI models due to space constraints, I will briefly outline two primary arguments defending the idea that DNNs can be meaningfully interpreted through a mechanistic approach.

First, DNNs maintain their mechanistic status through weak *structural constraints*, typical of functional explanations that require structural properties to realize a functional characterization (Piccinini, 2015). Piccinini and Craver (2011, p. 302) state, “The functional properties of black boxes are specified in terms of their inputs and outputs (plus the algorithm, perhaps), independently of their physical implementation.” DNNs exemplify this, as their inputs, outputs, and mapping algorithms can be defined without specific physical ties. However, the functional properties of DNNs impose certain constraints on the structural components regarding the degrees of freedom necessary for implementing an algorithm. This limits the functional analysis of an AI system to explaining how the algorithm, data, and network architecture determine the required degrees of freedom. If the system's functional components

are organized and can reliably differentiate between computational vehicles, the same computations can be implemented across various physical media—mechanical, electromechanical, electronic, or magnetic systems—without necessarily being affected by the specific properties of the physical medium.

Therefore, deep learning models can be viewed as mathematically defined systems that describe concrete, physically instantiated systems to some degree of approximation (cf. Piccinini, 2015). These models' functional properties specify the necessary degrees of freedom that a concrete system requires to perform computations. As the models are not causally complete enough to be considered typical mechanisms, their functional analyses can be described as "mechanism sketches," in which some structural aspects of a mechanistic explanation are intentionally omitted (Piccinini & Craver, 2011). This level of characterization remains relevant for XAI research, where low-level physical details are not always needed to determine the success or failure of algorithmic decision-making. An exception is when, for instance, the speed of information processing provided by the physical medium is crucial for avoiding miscomputation—for example when running a large language model on a smartphone or using an inadequate processing unit for image recognition in autonomous vehicles. However, case studies from OpenAI and Anthropic show that mechanistic decomposition of an AI system can focus on functional analysis without covering every aspect of computational phenomena in concrete processing systems.

Secondly, deep learning models can be seen as *teleofunctional mechanisms*, or simply *functional mechanisms*, which are defined by having teleological functions: specific "purposes" or "ends." Computing systems are generally designed with the teleological function of computing, which involves manipulating variables or specific values of variables according to rules sensitive to their properties (Piccinini, 2015; Coelho Mollo, 2018). In this context, deep learning systems perform digital computations by manipulating digits and strings of digits, which, although eventually translating to physical quantities—intervals of voltage values—are numerical vector representations of input data and internal states manipulated through matrix operations.

DNNs operate with medium-independent vehicles, which are the functional components of a mechanism (its entities), and the manipulations that these vehicles undergo (activities), according to mappings from inputs to outputs determined by a transition function set by the learning algorithm. These systems consist of organized components, each with specific functions, embodying the teleofunctional nature of DNN computation. When properly

organized and functioning, the coordinated activities of these components define the capabilities of DNN-style computation, provided that physical instantiation details offer the necessary degrees of freedom and do not lead to miscomputations. Assessing deep learning mechanisms to ascertain whether they fulfill their intended teleological functions involves decomposing the model into its components to understand their contributions. This extends to computing systems defined purely mathematically, which stand to concrete ones in roughly the same relation that the triangles of geometry stand to concrete triangular objects:

> A similar notion of functional mechanism applies to computing systems that are defined purely mathematically, such as (unimplemented) Turing machines. Turing machines consist of a tape divided into squares and a processing device. The tape and processing device are explicitly defined as spatiotemporal components. They have functions (storing letters; moving along the tape; reading, erasing, and writing letters on the tape) and an organization (the processing device moves along the tape one square at a time, etc.). Finally, the organized activities of their components explain the computations they perform (Piccinini, 2015, pp. 119–120).

Thus, decomposing functional mechanisms, even if the target system is defined mathematically, results in an elliptical mechanistic explanation of the system's capacities.

To summarize, the core concept of the mechanistic approach to deep learning is that despite possible skepticism regarding the ontic status of DNNs, we can still effectively utilize neomechanistic discovery strategies—decomposition, localization, and recomposition—to gain valuable insights into the internal workings of these systems. By pursuing mechanistic analysis augmented by case-specific analytical explainability techniques, researchers can identify functionally relevant components within the system and determine their precise roles, thereby obtaining EREs.

5 Epistemic Relevance of the Mechanistic Approach

While the relevance of the mechanistic approach in the XAI landscape has been addressed in existing technical literature and philosophical works, such as those by Kästner and Crook (2024), this paper advances the discourse by specifically focusing on the epistemic advantages and limitations of the mechanistic explanatory strategy and situating them within selected philosophical debates on XAI.

5.1 Epistemic Advantages

First, mechanistic decompositions of AI systems align with the interpretability criteria articulated by Doran et al. (2017). According to their framework, an *interpretable system* allows users not only to see, but also to study and understand how inputs are mathematically mapped to outputs. This is said to "help probe the mechanisms of ML systems." The authors cite regression, support vector machines, decision trees, ANOVAs, and data clustering models as

examples of interpretable systems. While acknowledging the interpretability challenges posed by DNNs, which autonomously learn and transform input features through nonlinear operations, mechanistic decomposition emerges as a promising method for examining their internal dynamics. By identifying human-understandable functional components, this approach aids technically proficient stakeholders in analyzing algorithmic mechanisms.

Second, mechanistic decomposition offers valuable insights into the internal mappings of AI systems, enhancing their structural transparency. Creel (2020) defines *structural transparency* as understanding how an algorithm is realized in code, involving knowledge of sub-components and their relationships, typically gained through analyzing system interactions. This understanding allows for modeling the system's structure and behavior using tools like code maps, flowcharts, and diagrams, closely aligning with mechanistic principles. Creel, however, is skeptical about the effectiveness of such decomposition in reducing the structural opacity of DNNs, noting that "when the functional units of the program are tiny, simple, and numerous, as are the neurons of a deep neural network, a subcomponent map would prove insufficient" (Creel, 2020, pp. 19–20). Nevertheless, recent advancements in XAI suggest that deeper insights into model functionality can be achieved through alternative units of analysis beyond individual neurons. Examples include neuron circuits (from OpenAI case study; Cammarata et al., 2021), and patterns of neuron activations (from Anthropic case study; Bricken et al., 2023a; 2023b), identified using mechanistic discovery strategies and techniques like feature visualization or dictionary learning. While the decomposition of complex neural networks into functionally relevant parts might obscure some neuron-specific operations, the mechanistic approach can significantly reduce structural opacity.

Third, the mechanistic explanatory strategy can yield highly *generalizable* and *counterfactual explanations* for AI decision-making. Drawing on Woodward's (2003) interventionist account of causality, Buijsman (2022) argues that effective AI explanations should demonstrate outcomes in counterfactual scenarios. Counterfactual descriptions compare a model's actual outcome with hypothetical alternatives to reveal input–output correlations. Buijsman suggests these contrasts are explanatory if they show generalizable correlations inferred from counterfactual reasoning, covering a range of "what-if-things-had-been-different" scenarios. Although he doesn't specifically analyze the mechanistic approach to XAI, mechanistic AI explanations can pinpoint some general rules governing model behavior across various models. This meets Buijsman's "reasonable generalizations" criterion, at least to the extent of addressing "what-if" questions.

Although mechanistic approaches typically focus on actual causal processes and may overlook potential counterfactual scenarios, Buijsman's perspective is based on Woodward's (2003) definition of causation, which involves "interventions": altering one variable without changing others that could affect the outcome. This defines counterfactual dependence as "*x* causes *y* if an intervention on *x* changes the value of *y*" (Buijsman, 2022, p. 566). The adoption of this view integrates causal-mechanistic interventions within "what-if" scenarios, suggesting that mechanistic considerations could fulfill Buijsman's criteria for counterfactual reasoning. Researchers can leverage typically mechanistic knowledge to explore such scenarios, for example, by perturbing input images to assess network resilience or by evaluating the impact of removing specific entities from the model.

Finally, Tomsett et al. (2018) argue that explainability of a machine learning system should be assessed relative to specific stakeholders or tasks. The black box problem arises because developers struggle to explain system behaviors through their learnable parameters; however, other stakeholders may seek different explanations to meet their needs (Zednik, 2019). *Stakeholders* range from developers responsible for engineering and maintaining the system to end users concerned with fairness, each requiring tailored explanations (Tomsett et al., 2018; Hind, 2019; Kasirzadeh, 2021).[3] In many cases, understanding an opaque system does not involve detailing parameter values, rather it involves comprehending the environmental patterns and abstract representational structures that the system models (Buckner, 2019; Zednik, 2019). This highlights the importance of recognizing diverse epistemic needs among stakeholders, especially regarding complexity and domain-specific language, to effectively fulfill their roles within an AI ecosystem.

The adoption of a mechanistic strategy for XAI, which emphasizes multilevel hierarchies of mechanisms, can address stakeholders' needs by offering understandable explanations at multiple levels. Mechanistic reasoning enables AI developers to use nested hierarchical structures to identify key causal variables, like learnable parameters and representational structures, that transform inputs into outputs. Consequently, detailed lower-level explanations can help developers identify errors or enhance performance, addressing their epistemic needs for model control, manipulation, and prediction. Conversely, based on empirical studies, Ribeiro et al. (2016) argue that while machine learning experts can navigate such complex landscapes, laypersons prefer explanations that reduce models to a small number of weighted

[3] Specific understanding of stakeholders is brought forward by the EU AI Act, which focuses on *deployers*: natural or legal persons, including a public authority, agency or other body, using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity (European Union, 2024).

features. Therefore, high-level, simplified descriptions of mechanisms and abstract schemata may better accommodate end users' epistemic and practical needs related to trust by distilling complex information into comprehensible knowledge.

This also aligns with Buijsman's (2022) call for abstract variables in explanations to increase generality and reduce cognitive load. A mechanistic multilevel approach may thus allow researchers to tailor explanations to diverse audiences, assisting also in meeting legal compliance requirements, such as those posed by the EU AI Act (European Union, 2024), by varying abstraction levels. Recall Anthropic's research on language model decomposition, which involved creating a "knob" to adjust the model's visibility resolution and experimenting with the number of features learned (Bricken et al., 2023a). This way, while the mechanistic explanatory strategy aims to provide a detailed understanding of the internal workings of AI models, it also offers a means for crafting human-grounded explanations that align with the epistemic requirements of various stakeholders.

Overall, to properly assess the impact of such multilevel approaches, more empirical studies involving multiple AI stakeholders are essential in gaining insights into the way explanations are perceived and understood by various audiences (cf., e.g., review of empirical studies on human-grounded explanations in Dorsch & Moll, Forthcoming).

5.2 Epistemic Limitations

While a theoretically grounded mechanistic exploratory strategy appears promising for the XAI program in terms of its epistemic advantages, its overall utility is also heavily constrained by certain epistemic limitations.

First, adopting Doran et al.'s (2017) XAI typology, mechanistic explanations enhance interpretability by revealing AI systems' inner workings, but they do not necessarily improve the comprehensibility of such system. *Comprehensibility* involves users making sense of outputs through interpretable symbols like words or visualizations, regardless of the system's internal opacity. "Auto-interpretability" techniques, such as those by Bricken et al. (2023b) and Schwettmann et al. (2023), generate natural language descriptions of model components to aid comprehension. Yet, understanding these symbols typically depends on the user's implicit knowledge. While visualization techniques might display recognizable features in image recognition systems, XAI methods often identify subtle, complex features that may elude human understanding (Buckner, 2019; Zednik, 2019). Thus, although explanations should

ideally present mechanisms in human-understandable terms, the statistical nature of deep learning frequently diverges from intuitive concepts.

Moreover, comprehending these explanations requires a certain level of technical proficiency in AI methods, which varies across types of explanations. Higher-level explanations that abstract away from intricate details generally require less expertise compared to detailed, lower-level explanations. This need for varying levels of expertise was evident in Ribeiro et al.'s (2016) evaluation of the LIME method, which specifically involved trained computer science graduates. This presents a significant challenge, as stakeholders beyond system developers will continue to demand explanations for system behavior, even when they lack the necessary technical background.

Second, Creel's (2020) distinction between types of transparency indicates that while mechanistic treatment may support structural transparency, it may fall short in achieving algorithmic and run transparency. *Algorithmic transparency* refers to knowledge of the algorithmic functioning of the whole, revealing high-level logical rules governing system transformations, which is not secured by mechanistic function-by-function decomposition. *Run transparency*, on the other hand, requires knowledge of specific program operations, including hardware specifics and input data. It involves observing how programs execute on particular hardware with real data. Since the mechanistic explanatory strategy presented here focuses on abstract models defined by mathematical constructs and specified degrees of freedom, it may not capture the artifacts of real-time interactions between software and hardware, unexpected data inputs, or the effects of software being converted into machine code.

Finally, there is the issue of *complexity*. While classical computer systems are relatively transparent, deep learning systems are considered black boxes due to the complex interdependencies among millions of parameters composing their internal states. This complexity enables neural networks to excel in problem solving but complicates the dissection of their causal–mechanical structure. Kostić (2023) points out that the opacity resulting from a model's functional complexity makes achieving a mechanistic explanation—requiring detailed knowledge of its components, activities, and organization—practically unattainable from the start. Even if not epistemically impossible, realistically addressing this challenge is practically daunting, given the limitations of current engineering methods, resources, and the increasing demand for explanations in rapidly evolving AI technologies. Perhaps the best we can hope for with the mechanistic approach is the examination of small, localized mechanisms, akin to that which occurs in neuroscience.

Complex DNNs often resemble *non-decomposable systems*, in which each component's behavior is heavily influenced by its interactions with many others (cf. Rathkopf, 2018). While decomposition helps in managing the complexity of representing every element, thereby mitigating combinatorial explosion, it often results in representations that are limited in scope and applicable mainly to specific subsystems or simplified toy models. Creel (2020) notes that while some input–output paths of a model might be straightforward, fully understanding all sub-components can be excessively complex. To address this challenge, there is growing interest in scaling microscopic insights from mechanistic interpretability research to broader understanding of larger models. [4] However, skepticism about such scalability persists due to computational challenges, high costs, and unresolved methodological questions (e.g., Nanda, 2023; Casper, 2023; Greenblatt et al., 2023). Evidence from small-scale models does not guarantee that real-world DNNs can be effectively decomposed for a thorough mechanistic understanding. When part–whole decomposition isn't feasible, alternative approaches that are fueled by a system's complexity, such as *network science* and *topological explanations* (e.g., Rathkopf, 2018; Kostić, 2022), should be considered.

6 Conclusions

The mechanistic explanatory strategy for XAI focuses on identifying the mechanisms that drive automated decision-making. In the case of deep neural networks, this requires discerning functionally relevant components—such as neurons, layers, circuits, or activation patterns—and understanding their exact roles through heuristic discovery strategies of decomposition, localization, and recomposition. Research suggests that such a coordinated, systematic approach to studying the functional organization of models can expose previously unrecognized elements that simple explainability techniques might miss, ultimately fostering more explainable AI. In this spirit, supported by real-world examples from image recognition and language modeling, this philosophical analysis underscores the value of mechanistic reasoning in XAI.

The mechanistic approach offers significant epistemic benefits, enhancing AI interpretability, structural transparency, and enabling crafting counterfactual and highly

[4] For instance, in OpenAI's research on curve detectors, researchers demonstrated how the first four layers of the InceptionV1 network gradually build towards curve detectors in the fifth layer, reverse engineering the operation of a family of 10 curve-detecting neurons (Cammarata et al., 2021). However, the complete InceptionV1 model consists of 22 layers (27 layers if counting pooling) with between 5 and 6 million parameters. Similarly, in an anthropic study on a transformer language model, researchers chose to examine a small, one-layer transformer with a 512-neuron layer (Bricken et al., 2023a). In comparison, GPT-3, the immediate predecessor of GPT-3.5 used in ChatGPT, features 175 billion parameters and operates within 96 layers.

generalizable explanations. This approach aids in prediction and system manipulation, as understanding internal dynamics allows for effective interventions and forecasting future states in new contexts. Additionally, it leverages multilevel hierarchical explanations, making complex AI systems more accessible and manageable for diverse stakeholders. Deepening our understanding of AI mechanisms and their causal relationships can improve performance evaluation and identify areas for improvement. Consequently, the advancement of a mechanistic framework in XAI may be crucial for overcoming trust and transparency challenges in high-stakes algorithmic decision-making.

However, despite its theoretical promise, the mechanistic strategy's utility faces significant epistemic limitations. These challenges include ambiguous effects on algorithmic and run transparency, the limited comprehensibility of explanations due to the lack of suitable concepts, the necessity for recipients to have technical proficiency in AI, and difficulties in decomposing complex, real-world systems, which are not evident in simpler toy model examples.

Several areas of future research stem from these considerations. Primarily, it is important to assess the applicability and scalability of the mechanistic approach beyond deep learning toy models to more complex AI systems. Further, investigating the way that individual stakeholders comprehend mechanistic AI explanations through user studies complimented by suitable epistemological theory of understanding would be a reasonable next step. Given the identified limitations of the mechanistic approach, it is also vital to explore other philosophically informed explanatory strategies. These should be ones that thrive on—rather than are hindered by—system complexity, and that also address other contexts of opacity.

Funding

This work was supported by the National Science Centre, Poland, under PRELUDIUM grant no. 2023/49/N/HS1/02461, and by the Polish National Agency for Academic Exchange under NAWA STER project no. BPI/STE/2021/1/00030/U/00001.

Acknowledgements

I would like to thank Marcin Miłkowski, Dimitri Coelho Mollo, Lena Kästner, Barnaby Crook, Kristian González Barman, and John Dorsch for their constructive comments that helped improve this manuscript.

References

Bechtel, W., & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421–441. https://doi.org/10.1016/j.shpsc.2005.03.010

Bechtel, W., & Abrahamsen, A. A. (2013). Thinking dynamically about biological mechanisms: Networks of coupled oscillators. *Foundations of Science*, 18, 707–723. https://doi.org/10.1007/s10699-012-9301-z

Bechtel, W., & Richardson, R.C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research. Second Edition.* Cambridge, MA: MIT Press/Bradford Books. https://doi.org/10.7551/mitpress/8328.001.0001

Beisbart, C., & Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*. https://doi.org/10.1111/phc3.12830

Bricken, T., Templeton, A., Batson, J., Olah, C., Henighan, T., Carter, S., Hume, T., Burke, J. E., McLean, B., Nguyen, K., Tamkin, A., Joseph, N., Maxwell, T., Schiefer, N., Kravec, S., Wu, Y., Lasenby, R., Askell, A., Denison, C., … Chen, B. (2023a, October 4). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, Anthropic. Retrieved from https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bricken, T., Templeton, A., Batson, J., Olah, C., Henighan, T., Carter, S., Hume, T., Burke, J. E., McLean, B., Nguyen, K., Tamkin, A., Joseph, N., Maxwell, T., Schiefer, N., Kravec, S., Wu, Y., Lasenby, R., Askell, A., Denison, C., … Chen, B. (2023b, October 5). Decomposing language models into understandable components. *Transformer Circuits Thread*, Anthropic. Retrieved from https://www.anthropic.com/index/decomposing-language-models-into-understandable-components

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10), e12625. https://doi.org/10.1111/phc3.12625

Buijsman, S. (2022). Defining explanation and explanatory depth in XAI. *Minds and Machines*, 32(3), 563–584. https://doi.org/10.1007/s11023-022-09607-9

Cammarata, N., Goh, G., Carter, S., Voss, C., Schubert, L., & Olah, C. (2021). *Curve circuits*. Distill, 6(1), e00024.006. https://doi.org/10.23915/distill.00024.006

Casper, S. (2023, February 17). EIS VI: Critiques of mechanistic interpretability work in AI safety. *AI Alignment Forum.* Retrieved from https://www.alignmentforum.org/posts/wt7HXaCWzuKQipqz3/eis-vi-critiques-of-mechanistic-interpretability-work-in-ai

Chan, L., Garriga-Alonso, A., Goldowsky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A., Shlegeris, B., & Thomas, N. (2022, December 3). Causal scrubbing: A method for rigorously testing interpretability hypotheses. *AI Alignment Forum.* Retrieved from https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing

Christiano, P. (2022, November 25). Mechanistic anomaly detection and ELK. *AI Alignment / Medium*. Retrieved from https://ai-alignment.com/mechanistic-anomaly-detection-and-elk-fb84f4c6d0dc

Coelho Mollo, D. (2018). Functional individuation, mechanistic implementation: the proper way of seeing the mechanistic view of concrete computation. *Synthese*, 195, 3477–3497. https://doi.org/10.1007/s11229-017-1380-5

Conmy, A., Mavor-Parker, A.N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. *ArXiv*, abs/2304.14997. https://doi.org/10.48550/arXiv.2304.14997

Craver, C.F. (2007). Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience, Oxford: Clarendon Press.

Creel, K. A. (2020). Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4), 568–589. https://doi.org/10.1086/709729

Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600. https://doi.org/10.48550/arXiv.2309.08600

Doran, D., Schulz, S.C. & Besold, T.R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. In *CEUR Workshop Proceedings*, 2071. https://doi.org/10.48550/arXiv.1710.00794

Dorsch, J., & Moll, M. (Forthcoming). Explainable and Human-Grounded AI for Decision Support Systems: The Theory of Epistemic Quasi-Partnerships. In Müller, V. C., Dewey, A. R., Dung, L., & Löhr, G. (Eds.), *Philosophy of Artificial Intelligence: The State of the Art*. Synthese Library, Berlin: Springer Nature. https://doi.org/10.48550/arXiv.2409.14839

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., ... & Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Retrieved from https://transformer-circuits.pub/2021/framework/index.html

Erasmus, A., Brunet, T.D.P., & Fisher, E. (2021). What is Interpretability? *Philosophy & Technology*, 34, 833–862. https://doi.org/10.1007/s13347-020-00435-2

European Union. (2024). *EU Artificial Intelligence Act*. Retrieved from https://artificialintelligenceact.eu

Glennan, S.S. (1996). Mechanisms and The Nature of Causation. *Erkenntnis*, 44, 49–71. https://doi.org/10.1007/BF00172853

Glennan, S.S. (2017). *The New Mechanical Philosophy*. Oxford: Oxford University Press.

Greenblatt, R., Nanda, N., Buck, Shlegeris, B., Habryka, O. (2023, December 1). How useful is mechanistic interpretability? *Lesswrong.* Retrieved from https://www.lesswrong.com/posts/tEPHGZAb63dfq2v8n/how-useful-is-mechanistic-interpretability

Haimovici, S. (2013). A problem for the mechanistic account of computation. *Journal of Cognitive Science*, 14(2), 151–181. http://doi.org/10.17791/jcs.2013.14.2.151

Hind, M. (2019). Explaining Explainable AI. *XRDS: Crossroads, The ACM Magazine for Students*, 25, 16–19. https://doi.org/10.1145/3313096

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Illari, P., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, 2(1), 119–135. http://dx.doi.org/10.1007/s13194-011-0038-2

Karmakharm, T. (2018). Image classification with DIGITS. *NVIDIA Deep Learning Institute.* Retrieved from https://rse.shef.ac.uk/assets/slides/2018-07-19-dl-cv/image-classification.pdf

Kasirzadeh, A. (2021). Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, Association for Computing Machinery, New York, NY, USA, 14. https://doi.org/10.1145/3442188.3445866

Kästner, L., & Crook, B. (2024). Explaining AI through mechanistic interpretability. *European Journal for Philosophy of Science*, 14, 52. https://doi.org/10.1007/s13194-024-00614-4

Kostić, D. (2022). Topological explanations: An opinionated appraisal. In I. Lawler, E. Shech, & K. Khalifa (Eds.), *Scientific Understanding and Representation: Mathematical Modeling in the Life and Physical Sciences* (pp. 96–115). Routledge. https://doi.org/10.4324/9781003202905-9

Kostić, D. (2023). *Pragmatics of Explainability Relevance in XAI*. Manuscript.

Kostić, D., & Halffman, W. (2023). Mapping explanatory language in neuroscience. *Synthese*, 202, 112. https://doi.org/10.1007/s11229-023-04329-6

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S.B. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 18. https://doi.org/10.3390/e23010018

Lipton, Z.C. (2018). The Mythos of Model Interpretability. In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Machamer, P.K., Darden, L., & Craver, C.F. (2000). Thinking about Mechanisms. *Philosophy of Science*, 67(1), 1–25. https://doi.org/10.1086/392759

Miłkowski, M. (2014). Computational mechanisms and models of computation. *Philosophia Scientiæ*, 18-3, 215–228. https://doi.org/10.4000/philosophiascientiae.1019

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T., Howe, P.D., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *ArXiv*, abs/1712.00547. https://doi.org/10.48550/arXiv.1712.00547

Mittelstadt, B.D., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT* '19),* Association for Computing Machinery, New York, NY, USA, 279–288. https://dl.acm.org/doi/10.1145/3287560.3287574

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book/

Montavon, G., Samek, W., & Müller, K.R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Nanda, N. (2023, July 6). Concrete open problems in mechanistic interpretability: A technical overview. *Effective Altruism Forum*. Retrieved from https://forum.effectivealtruism.org/posts/EMfLZXvwiEioPWPga/concrete-open-problems-in-mechanistic-interpretability-a

Olah, C. (2022). Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuit Thread*, OpenAI. Retrieved from. https://transformer-circuits.pub/2022/mech-interp-essay/index.html

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*. https://doi.org/10.23915/distill.00007
Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T.J., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., … Olah, C. (2022). In-context Learning and Induction Heads. *ArXiv*, abs/2209.11895. https://doi.org/10.48550/arXiv.2209.11895
Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, 29, 441–459. https://doi.org/10.1007/s11023-019-09502-w
Piccinini, G. (2015). *Physical Computation: A Mechanistic Account.* Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199658855.001.0001
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese* 183, 283–311. https://doi.org/10.1007/s11229-011-9898-4
Qin, Z., Yu, F., Liu, C., & Chen, X. (2018). How convolutional neural network see the world — A survey of convolutional neural network visualization methods. *Mathematical Foundations of Computing*, 1(2), 149–180. https://doi.org/10.3934/mfc.2018008
Rathkopf, C. (2018) Network representation and complex systems. *Synthese* 195, 55–78. https://doi.org/10.1007/s11229-015-0726-0
Ribeiro, M., 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*. https://doi.org/10.1145/2939672.2939778
Schuman, C.D., Kulkarni, S.R., Parsa, M. et al. Opportunities for neuromorphic computing algorithms and applications. *Nat Comput Sci*, 2, 10–19 (2022). https://doi.org/10.1038/s43588-021-00184-y
Schwettmann, S., Shaham, T.R., Materzynska, J., Chowdhury, N., Li, S., Andreas, J., Bau, D., & Torralba, A. (2023). FIND: A function description benchmark for evaluating interpretability methods. *ArXiv*, abs/2309.03886. https://doi.org/10.48550/arXiv.2309.03886
Schyns, P.G., Snoek, L., & Daube, C. (2022). Degrees of algorithmic equivalence between the brain and its DNN models. *Trends in Cognitive Sciences*, 26, 1090–1102. https://doi.org/10.1016/j.tics.2022.09.003
Shahriar, N. (2023, February 1) What is Convolutional Neural Network — CNN (Deep Learning). Retrieved from https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5
Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *ArXiv*, abs/1806.07552. https://doi.org/10.48550/arXiv.1806.07552
Watson, D.S., & Floridi, L. (2020). The explanation game: A formal framework for interpretable machine learning. *Synthese*, 198 (10), 9211–9242. https://doi.org/10.1007/s11229-020-02629-9
Woodward, J. (2003). Making Things Happen: A Theory of Causal Explanation. Oxford University Press.
Wright, C., & Bechtel, W. (2007). Mechanisms and psychological explanation. In Thagard, P. (Ed.), *Philosophy of psychology and cognitive science*, Elsevier.
Yosinski, J., Clune, J., Nguyen, A.M., Fuchs, T.J., & Lipson, H. (2015). Understanding neural networks through deep visualization. *ArXiv*, abs/1506.06579. https://doi.org/10.48550/arXiv.1506.06579
Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology*. 34 (2), 265–288. https://doi.org/10.1007/s13347-019-00382-7
Zerilli, J., Knott, A., MacLaurin, J., & Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, 32, 661–683. https://doi.org/10.1007/s13347-018-0330-6