European Journal for Philosophy of Science (2024) 14:52
https://doi.org/10.1007/s13194-024-00614-4

PAPER IN PHILOSOPHY OF TECHNOLOGY AND INFORMATION

Explaining AI through mechanistic interpretability

Lena Kästner[1] · **Barnaby Crook**[1]

Received: 29 January 2024 / Accepted: 12 September 2024 / Published online: 11 October 2024

Abstract
Recent work in explainable artificial intelligence (XAI) attempts to render opaque AI systems understandable through a divide-and-conquer strategy. However, this fails to illuminate *how trained AI systems work as a whole*. Precisely this kind of functional understanding is needed, though, to satisfy important societal desiderata such as safety. To remedy this situation, we argue, AI researchers should seek *mechanistic interpretability*, viz. apply coordinated discovery strategies familiar from the life sciences to uncover the functional organisation of complex AI systems. Additionally, theorists should accommodate for the unique costs and benefits of such strategies in their portrayals of XAI research.

Keywords AI · ANN · Deep learning · Discovery · Explanation · Mechanistic interpretability · XAI

1 Introduction

Over the past decade, the term "AI" has increasingly become a synonym for deep artificial neural networks (ANNs) trained with machine learning (ML) algorithms. These ANNs are often complex and opaque, approximating target functions through the mutual contribution of millions or even billions of parameters with values learned during an automated training process (Baraniuk et al., 2020; Chollet, 2021; Russell & Norvig, 2020). On the one hand, this setup allows ANNs to exhibit flexible and expressive behaviour, developing sophisticated ways of representing and processing information. On the other hand, these systems can be surprisingly brittle and prone to unpredictable and catastrophic failure (Fawzi et al., 2018; Raghu et al., 2017). Since the functional organisation of ANNs is both complex and machine-learned, *how* their internal structure implements the mapping from inputs to outputs

✉ Lena Kästner
lena.kaestner@uni-bayreuth.de

Barnaby Crook
barnaby.cook@uni-bayreuth.de

[1] Department of Philosophy, University of Bayreuth, Bayreuth, Germany

often remains unknown. Despite this, AI systems based on ANNs are increasingly used to support or take over human tasks – from making music recommendations to medical decisions (Huang et al., 2020; Jugovac & Jannach, 2017). Especially in high-stakes[1] circumstances, regulators, philosophers and AI researchers increasingly call for making AI systems *explainable* (Burrell, 2016; Zednik, 2021).

To address this need, a research field has developed under the heading of explainable AI (XAI). But far from being unified, research under this umbrella tackles a range of distinct, though interrelated issues: Computer scientists take on the technical challenge of developing computational methods to produce explanations of AI systems (Guidotti et al., 2018; Ribeiro et al., 2016; Wachter et al., 2017). Philosophers attempt to characterise relevant societal desiderata such as safety, trustworthiness, reliability, and fairness (Durán & Jongsma, 2021; Langer et al., 2021a, 2021b; Páez, 2019). And psychologists grapple with the problem of quantifying epistemic outcomes like *understanding* (Langer et al., 2021a, 2021b; Sloman & Rabb, 2016). In the face of this complexity, scholars have tried to coordinate these different strands of research, highlight how they might be mutually advantageous, and assess the extent to which (and at what costs) XAI can achieve its goals. To this end, XAI researchers have developed taxonomies of explainability methods (Nunes & Jannach, 2017; Speith, 2022), offered conceptual models of the so-called explainability problem (Langer et al., 2021a, 2021b; Zednik, 2021), and proposed simplifying unifications (Fleisher, 2022; Nyrup & Robinson, 2022). We applaud these contributions, and agree that AI explainability is a multifaceted issue which resists any one-size-fits-all solution (cf., Langer et al., 2021a, 2021b; Zednik, 2021). However, the picture of XAI research that is typically presented is one in which computational methods deliver explanations for specific stakeholders in particular contexts. We think this is crucially incomplete: we must also consider the potential of coordinated research strategies aiming to uncover the functional organisation of trained AI systems.

While research to this end is currently gaining momentum among researchers working in industry (e.g., Cammarata et al., 2020; Olah et al., 2018), attention to it is still limited within the academic sphere (but see, e.g., Bau et al., 2017; Geiger et al., 2022, 2023). Contemporary XAI effectively pursues a *divide-and-conquer strategy* by focusing on how individual methods might deliver explainability in specific contexts. While this strategy can successfully highlight which features were influential for a given outcome, or how input features would need to be changed to obtain a different decision, this is only a small part of the story if we are seeking to explain and understand how AI systems work. Indeed, what we *should* aim for to ensure that AI systems fulfil desiderata commonly demanded of AI systems by society – such as safety – goes far beyond specific contexts: We need generalisable insights about *how* the systems in question *work as a whole*; for only such insights about the functional organisation that elicits behavioural patterns and dispositions will allow us to

[1] We refrain from committing to any particular definition of 'high-stakes'. See the European Commission's Artificial Intelligence Act (Artificial Intelligence Act, 2024) for extended discussion of what constitutes a 'high-risk' application of AI.

anticipate how systems might respond to novel inputs and how they might behave when exposed to yet unexplored contexts. Importantly, what is of interest here is not the architecture that an AI system's programmers have specified prior to training.[2] Rather, we are interested in the learned structure that emerges through the automated training procedures a system undergoes after its initial specification; hence it is also referred to as *emergent structure* (Manning et al., 2020). Crucially, we argue, these insights cannot usually be extrapolated from applying individual XAI methods tailored to specific contexts – we would simply be missing the forest for the trees.

We propose that to remedy this situation, XAI researchers should take a *mechanistic interpretability (MI) approach* for complex AI systems. In fact, research in line with the approach we suggest is already being pursued by a small community of researchers (Section 3.2). It starts from the premise that, once AI systems become sufficiently complex, they are best investigated and explained through the same lens as biological organisms (rather than being treated as technical artefacts). Thus, practitioners should seek to characterise AI systems in terms of their functional organisation (the organised activities of their functionally relevant components).[3] This requires the application of coordinated discovery strategies familiar from life sciences, such as pattern recognition, functional decomposition, localisation, and systematic experimental manipulations. As such, MI research may be significantly more resource-intensive (both in terms of time and in terms of labour) than the divide-and-conquer strategy. In return for this investment, we gain a deeper and more holistic understanding of how trained AI systems work. We argue that the outcome of effective MI research affords greater predictivity and, by enabling surgical interventions, greater control of system behaviour. Thus, we claim, MI enables us to meet important societal desiderata including safety. Given these distinctive costs and benefits, and the fact that the MI approach has been successfully employed by a small community of researchers, we think it is crucial for XAI theorists (as well as philosophers of science more generally) to accommodate this research strategy in their analyses of the field. For only by having the full range of strategies for explaining complex AI systems in view can we make appropriate choices about which research to pursue, and thereby ensure that crucial desiderata will actually be satisfied when employing opaque AI systems in society.

We shall proceed as follows: In Section 2, we briefly discuss achievements and limitations of contemporary XAI research. In Section 3, we propose that rather than adopting a divide-and-conquer strategy, XAI researchers should seek explainability through mechanistic interpretability. We illustrate this approach by examining the case of Distill. In Section 4, we discuss possible objections to our proposal. In Section 5, we conclude.

[2] For ANNs, the number of layers, number of units (neurons) per layer, activation functions of units, connectivity of units, loss function, and learning algorithm (including hyperparameters like batch size and learning rate) are all typically pre-programmed.

[3] Functionally relevant components can be any unit or structure that serves to achieve a specific function within a system; this is explicitly not limited to pre-specified entities like neurons but may include complex structures such as, e.g., circuits (see Section 3.1) or distributed representations.

2 Contemporary XAI: The divide-and-conquer strategy

For current purposes we focus on ML-based AI systems (with ANNs being a paradigm case) that are opaque due to the scale and complexity of their learned structure (Burrell, 2016). Following Humphreys (2009), we characterise opacity as a lack of knowledge about a system's epistemically relevant elements (EREs) (see also Zednik, 2021).[4] The term ERE is deliberately non-specific. It captures any robust patterns which underlie or maintain a system's behaviour and are relevant to the epistemic goals of an agent. Thus, a system's opacity is relative both to an agent's (e.g., a company, AI user, or developer) interests and their knowledge about the system at a given time. Against this backdrop, we assume that any information eliminating opacity (by uncovering EREs) can function as an explanation (cf., Nyrup & Robinson, 2022; Zednik, 2021). We take making AI systems explainable by reducing their opacity to be the goal of XAI.[5]

Contemporary XAI research tries to achieve this goal through technical means. Specifically, it aims to develop algorithmic procedures – XAI *methods* – that generate explanatory information about AI systems. Hence, XAI is very much "in the business of developing analytic techniques with which to render opaque computing systems transparent" (Zednik, 2021, p. 285; see Mittelstadt et al., 2019; Rudin, 2019 for similar claims). Since the kind of explanatory information needed to eliminate opacity depends on numerous factors (Kirsch, 2017; Langer et al., 2021a, 2021b), XAI researchers have developed a range of methods with different properties (e.g., scope, detail, format) (Guidotti et al., 2018; Molnar, 2022; Speith, 2022). Further work has addressed which methods are best suited for concrete scenarios in which stakeholders interact with AI systems (Barredo Arrieta et al., 2020; Belle & Papantonis, 2021). This approach assumes that rendering AI systems explainable requires i) developing a range of XAI methods, ii) identifying specific explainability contexts, and iii) mapping XAI methods to these contexts. The result is a *divide-and-conquer strategy*: XAI research seeks, in any given case, to provide "the most appropriate explanation for a specific ML solution in a given context on a given task" (Zhou et al., 2021, p. 2; see also Fleisher, 2022, p. 12).

The divide-and-conquer strategy has its merits. It naturally accommodates the insight that there cannot be a one-size-fits all XAI approach, and that different stakeholders, contexts and AI systems pose different constraints on feasible XAI methods. Further, the divide-and-conquer strategy has been and continues to be productive, both in terms of novel methods and conceptual analysis (e.g., Barredo Arrieta et al., 2020; Belle & Papantonis, 2021; Köhl et al., 2019; Langer et al., 2021a, 2021b; Ribeiro et al., 2016; Sokol & Flach, 2020; Zhou et al., 2021). Indeed, contemporary

[4] Notice that while we adopt Humphreys' framing in terms of *knowledge* of EREs, this can be adapted to an account centred on *understanding* by defining opacity as the lack of a *grasp* of EREs. Grasping an ERE involves making use of it to perform inferences, take decisions, or complete downstream tasks (Keil, 2019; Strevens, 2013).

[5] Achieving this outcome does not require comprehensive explanation of *all* AI systems. For systems that operate in low-stakes environments, fewer elements of the system may be deemed epistemically relevant.

XAI research successfully produces tools enabling stakeholders to understand narrow aspects of AI systems (Belle & Papantonis, 2021; Guidotti et al., 2018; Lapuschkin et al., 2019; Molnar, 2022; Ribeiro et al., 2016, 2018; Wachter et al., 2017). One prominent example is *Local Interpretable Model-Agnostic Explanations* (LIME) (Ribeiro et al., 2016). This technique, applicable to any AI classification system, produces linear surrogate models which highlight the importance of particular features for a given classification. LIME has been shown to improve stakeholders' ability to assess the quality of rival classifiers and, as such, may prove an adequate solution in specific contexts where such skills are required. Another success story for the divide-and-conquer strategy is research on *counterfactual explanations* and *algorithmic recourse* (Karimi et al., 2022; Wachter et al., 2017). This family of XAI methods can inform subjects of algorithmic decisions about which alterations to the input data would have resulted in a desired outcome. Such methods are crucial to enabling agency in an increasingly algorithmic world (Vredenburgh, 2022).

Despite these successes, however, the *divide-and-conquer strategy* is also limited. Neither LIME nor counterfactual explanations can provide a comprehensive understanding of *how the systems* under investigation *work as a whole*. They do not tell us *how exactly* the internal structure of the model maps inputs to outputs, neither do they allow us to predict how a system will behave in *novel contexts*. Indeed, in an influential paper, Rudin criticises the enterprise of XAI for this reason, asserting that XAI methods "do not provide enough detail to understand what the black box is doing" (2019, p. 208). Similarly, Freiesleben (2024) evaluates two specific XAI methods (*feature visualisation* and *network dissection*) and concludes that neither are sufficient to support inferences about the functional role of subparts of model structure. Importantly, this information cannot be inferred even by independent application of various XAI method outputs. What is needed instead is systematic investigation that yields generalisable insights about systems' overall functional organisation.

The need for such insights is expressed vividly in current debates about AI regulation in politics and society. Take the European Commission's proposed regulatory framework for AI systems (Artificial Intelligence Act, 2024). Article 14 states that adequate human oversight in high-risk scenarios demands that an overseer should "properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, also in view of detecting and addressing anomalies, dysfunctions and unexpected performance" and be able "to intervene on the operation of the high-risk AI system" (Artificial Intelligence Act, 2024, art. 14). With regards to both requirements, a detailed (and generalisable) understanding of a trained system's overall functional organisation is crucial. For illustration, consider the phenomenon of *typographic adversarial examples* in the image model *CLIP* (Goh et al., 2021). CLIP was trained to predict which text was paired with a given image on the internet. However, researchers found that CLIP was easily fooled by simply sticking a written label (e.g., "phone") on an object (e.g., an apple), causing the image to be classified according to the written label rather than the object (e.g., as phone rather than apple). This behaviour was not foreseeable from ordinary reliability testing. Neither does highlighting the paper label as relevant to the classification output explain *why* CLIP relied on the labels rather than the objects in the

images for classification. However, through systematic investigation into the model, researchers revealed high-level neurons in CLIP which are sensitive to both images and text (Goh et al., 2021). Understanding this feature of the model's functional organisation made it easy to explain and predict CLIP's behaviour when faced with written labels in images.

Luckily, CLIP is not in use in high-stakes domains. However, consider the dangers of deploying a system with similar vulnerabilities to, e.g., scan for weapons at an airport. Here, as in many other cases where AI is deployed in the real world, safety is among the chief desiderata. It is listed among the first objectives in the EU AI Act (Artificial Intelligence Act, 2024, art. 1).[6] One way to ensure safety and prevent harm is to subject AI systems to rigorous testing before deploying them in high-stakes situations (Durán & Jongsma, 2021). However, doing this exhaustively will often be infeasible due to the dimensionality of the input space (Hacker et al., 2023; Keogh & Mueen, 2017). This makes it difficult to rule out the possibility of unpredictable failure when systems encounter novel inputs (Amodei et al., 2016; Hendrycks et al., 2018, 2023; Wei et al., 2023). A necessary complementary strategy, we think, is to characterise how the systems in question function in terms of how relevant components work together to elicit the behaviour we observe, i.e., MI. Based on such characterisations, we can anticipate and interpret system behaviour even for novel situations, helping us to ensure system reliability and safety (Meyes et al., 2019). The case of CLIP goes to show that insights into a system's functional organisation can help us understand a system's capabilities and limitations and avoid unforeseen failures.

Another benefit of MI is enabling more precise and informed scientific communication about the nature and properties of opaque AI systems. The high-profile release of strikingly performant models such as ChatGPT has led to a surge of debate about these systems (M. Mitchell & Krakauer, 2023; Pavlick, 2023). However, whether or not it is reasonable to describe large language models (LLMs) such as ChatGPT as possessing knowledge (Lam, 2022), implementing symbolic reasoning (Pavlick, 2023), or representing the world (Li et al., 2023) arguably depends upon empirical questions about the internal structure that implements the model's behaviour (Millière & Buckner, 2024). In other words, an accurate and scientifically defensible description of AI systems is contingent on the characterisation of EREs that may only be achieved through MI. We discuss the advantages of MI further in Section 3.3.

Importantly, the issue we are getting at here is a principled one, not a criticism of any particular XAI method. We acknowledge that certain methods are sufficient for explaining particular instances of system behaviour. But when deploying AI systems in the real world, fulfilling important societal desiderata requires a kind of *holistic*

[6] While there is no domain-general definition of what it means for an AI system to be safe, safety is typically associated with the avoidance of both physical and psychological harm to human beings (Steimers & Schneider, 2022).

explainability that provides generalisable insights into how a system's functional organisation elicits its behaviours or outputs.[7]

For clarity, we stress that the holistic explanations MI aims for are not synonymous with *global* explanations. In the XAI literature, *global* is an adjective ascribed to XAI methods which produce explanatory information about whole models (Molnar, 2022; Speith, 2022); it is contrasted with *local* methods which produce explanatory information about individual model inferences (e.g., one prediction or classification). *Holistic*, in contrast to both of these terms, is an adjective we use to refer to explanations (such explanations typically involve diagrams and accompanying text) which capture the *entire functional structure* of trained models (as MI aims to do). In principle, an idealised global XAI method *could* produce a holistic explanation. However, in practice, no extant method comes close to doing this. For example, Molnar (2022) categorises *partial dependence plots* as global methods, but plotting the marginal effect of particular features on model outputs is clearly a far cry from providing a holistic explanation of a trained model's functional structure.

Ultimately, we claim that the project of developing holistic explanations of sufficiently complex AI systems will *never* be adequately achieved through employing specific XAI methods in particular contexts. Instead, we propose, it can be achieved by *mechanistic interpretability* (MI). That is, by developing (human-interpretable) *mechanistic* explanations of systems' functional organisation. So far, this kind of project has been underappreciated by both philosophers and computer scientists.

3 Augmenting contemporary XAI through mechanistic interpretability

We suggest taking a *mechanistic interpretability (MI) approach* to complex AI systems that starts from the following premise: once AI systems become sufficiently complex, they are best investigated and explained through the same lens as biological organisms rather than being treated merely as technical artefacts (cf., Eden, 2007). Taking inspiration from successful scientific inquiry in the life sciences, we propose that XAI practitioners should apply coordinated discovery strategies (such as pattern recognition, functional decomposition, localisation, and systematic experimental manipulations) to characterise AI systems in terms of their functional organisation, i.e., the organised activities of their functionally relevant components. At the same time, theorists should accommodate the unique costs and benefits of this research strategy in their portrayals of XAI research. With the full range of strategies for explaining AI systems in view, we can make appropriate choices about how to ensure that crucial desiderata will be satisfied when deploying trained AI systems in society.

[7] We use the term 'holistic' to signal a focus on how AI systems' functional organisation implements their behaviour. As we will show, this need not be opposed to reductionist methodologies (see also Burnston, 2021).

3.1 Mechanistic interpretability

If we want to explain how AI systems work *as a whole*, we are essentially interested in their *functional organisation* or structure. That is, we seek to understand what system properties support their behaviour and how system functions are implemented by the orchestrated interactions of relevant component parts. This kind of project is neither new nor unique to XAI research.[8] And it is highly familiar from recent discussions on how to explain biological systems mechanistically (see Bechtel, 2009; Bechtel & Richardson, 1993; Craver, 2001; Craver & Darden, 2013). When mechanistic inquiry delivers explanations of sufficiently high quality, it affords improved prediction and control of target systems (Baetu, 2011; Craver, 2007; Howick et al., 2010; Winning & Bechtel, 2018; Woodward, 2017; Zou et al., 2023). Therefore, we suggest adopting strategies familiar from mechanistic inquiry in the life sciences to develop mechanistic explanations of the capabilities, limitations, and behaviours of opaque AI systems.

Recall that ANNs acquire their functional organisation through automated training procedures. After training, components within the system (the EREs of mechanistic explanations) will adopt specialised roles that programmers do not usually anticipate. These components will be organised in ways allowing them to work together to elicit the behaviours we observe (Manning et al., 2020; Richards et al., 2019). The task for researchers seeking MI will be to uncover the relevant components along with their organisation. This task may not be simple, and it may seem impossible to map functions onto discrete parts of an ANN at first sight (see Section 4 for further discussion). In fact, identifying relevant components will often go beyond investigating the functional roles of pre-individuated ANN structures like neurons and layers (Bau et al., 2017). Frequently, it may involve characterising more exotic and distributed structures like circuits (see Section 3.2), high-level representations (Zou et al., 2023), and representational subspaces (see Elhage et al., 2021).

To achieve such characterisations of functional components (EREs) in AI systems, researchers must detect and describe the robust patterns that underlie or maintain the system's behaviour (Wimsatt, 1994). In other words, they must engage in a *pattern recognition practice* (Haugeland, 1998; Kästner & Haueis, 2021), where patterns are any non-random arrangements within systems (Dennett, 1991) which (in virtue of their orderly character) serve as candidates for recognition. As such, patterns constitute potential EREs when seeking MI for AI systems.

Pattern recognition practices in science are collective endeavours involving the coordinated application of shared skills, tools, and concepts (Brigandt, 2011; Kästner & Haueis, 2021). They consist in a set of *epistemic activities* which conform to epistemic norms shared and continuously refined by the research community throughout scientific inquiry. Epistemic activities include such things as *decomposition* and *localisation* (cf. Bechtel & Richardson, 1993; Brigandt, 2011;

[8] The idea that certain aspects of computer science should be treated as empirical inquiry (as opposed to a branch of mathematics) goes back at least as far as Newell and Simon (1976), see Eden (2007) for discussion.

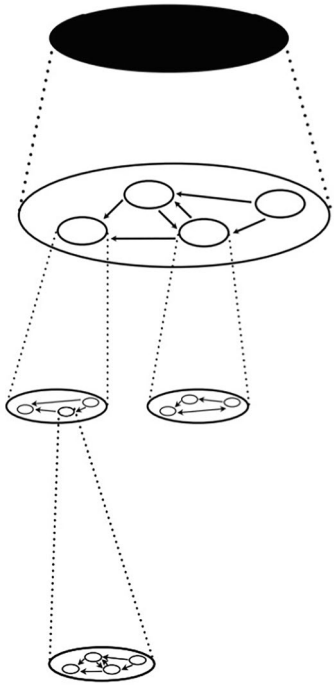

Fig. 1 The application of coordinated discovery strategies uncovers the functional organisation of trained AI systems. The system's overall behaviour (black circle at the top) is elicited by the relevant components (smaller circles) working together. Components can be further analysed into subcomponents to reveal a nested hierarchy (see also Craver, 2007)

Kästner & Haueis, 2021) or *recomposition* (cf. Bechtel & Abrahamsen, 2005). *Decomposition* involves breaking a phenomenon down into a set of constitutive sub-functions and is often applied recursively to produce multi-level functional characterisations (Craver, 2007). *Localisation* is the assignment of a sub-function to a particular part of the system which is hypothesised to implement it. *Recomposition* involves building a system back up in a functionally informed manner. To carry out an epistemic activity, researchers will typically engage in a variety of *epistemic operations*. These are the atomic units of scientific inquiry; they consist of concrete actions that track, measure, or manipulate the components of the target system, or generate, transform, and visualise data. In practice, epistemic activities are typically applied in an iterative manner to create a growing store of collective knowledge and fine-tune hypotheses about which system components are functionally relevant, how they are organised, and how they interact to produce characteristic phenomena.

As such, we argue, pattern recognition practices are ideally suited to produce interpretable, functional characterisations of *how AI systems work as a whole*. They deliver *mechanistic explanations* (Bechtel & Abrahamsen, 2005; Craver, 2001; Kästner & Haueis, 2021) of how the functionally relevant components of a system interact to produce a phenomenon of interest. In the best case, these explanations are comprehensive, providing a detailed, multi-level account of how a system works (see Fig. 1). This holds for artificial systems just as for biological ones.

3.2 The case of distill

Before turning to a case study, we note that MI has gained significant traction recently, with much work focused on transformer-based LLMs (Bricken et al., 2023; Li et al., 2023; Manning et al., 2020; Meng et al., 2023). One notable approach here is *dictionary learning*, in which sparse autoencoders are used to discover features which are represented across numerous neurons (Bricken et al., 2023; Rajamanoharan et al., 2024). Another particularly promising approach (not specific to transformers) involves causal abstractions, which enable hypotheses about the functional role of subgraphs of neural networks to be evaluated rigorously through systematic manipulations (Geiger et al., 2022, 2023). Yet another strategy, explicitly inspired by cognitive science, is *representation engineering*, which attempts to locate high-level concepts and processes within trained models (Zou et al., 2023).[9] Further avenues of active investigation within MI include studying the global representational properties of ANNs (Elhage et al., 2022), how mechanistic structure evolves throughout the learning trajectory (Nanda et al., 2023), and the potential automation of numerous aspects of the MI research process (Conmy et al., 2023; Hernandez et al., 2022; Marks et al., 2024). For useful surveys and discussion of the strengths and weaknesses of some of these methods see Räz (2023) and Räuker and colleagues (2023).[10] At this point, we wish to reiterate that while certain methods are particularly well-suited for MI, what distinguishes MI from traditional XAI is not the methods it uses, but the way it applies methods in coordination with scientific reasoning in pattern recognition practices that are systematic, coordinated, and iterative. As such, for brevity and clarity of exposition of this central point, we limit our philosophical analysis in this paper to the case of Distill.

In their work on the image classification ANN *InceptionV1* (Cammarata et al., 2020; Olah et al., 2018), Chris Olah and colleagues explicitly endorse applying life science research strategies to ANNs, suggesting that "neural networks are an object of empirical investigation" (Olah et al., 2020a). Hence, they set out to characterise the functional structure of InceptionV1 using coordinated discovery strategies as described above.

For reasons of space, we cannot describe the InceptionV1 architecture in detail here; but a brief introduction to the structure of convolutional ANNs (CNNs) will aid understanding of what follows. As this family of models was designed to process image data, their architecture reflects an attempt to match the compositional and hierarchical properties of natural images (Chollet, 2021). The convolutional layers that give CNNs their name consist of filters which

[9] Note that Zou and colleagues frame their approach as an *alternative* to MI, which they construe narrowly as the decomposition of ANNs into circuits. However, on our broader construal of MI as a pattern recognition practice, their approach clearly fits within it.

[10] The discussion of robust concept detection in Räz (2023, p. 5) is particularly pertinent to our discussion. However, we do not think the functional structure of trained AI systems necessarily needs to be explained in terms of concepts (see Boge, 2023 for discussion of whether ANNs learn concepts).

cover a small portion of the input image (represented as an array of raw pixel values). These filters slide over the image, computing activations (dot products of input and filter plus bias) at each spatial position. By analogy with biological neural networks (and mathematical equivalence with other ANN architectures), filters can be thought of as the receptive fields of neurons, with one neuron corresponding to each spatial position the filter covers.[11] As filters consist of trainable parameters, neurons can learn to respond to different features in the image, producing maps reflecting the presence or absence of features (e.g., edges) to feed to downstream layers. Through composition of simple features, neurons in these downstream layers can thus become sensitive to complex features like shapes, objects, and people. Thus, the very structure of CNNs (their depth, connectivity, and layer design) are adapted to the hierarchical compositionality that characterises images.

Olah and colleagues (Cammarata et al., 2021; Olah et al., 2020a) discover and analyse a *curve detector circuit* within InceptionV1. Circuits are sub-graphs of neural networks which, crucially, are not specified as distinct parts of the ANN's architecture. Instead, circuits are part of the model's learned structure. They are functional units which neurons self-organise into during the training process. As such, the curve detector circuit, consisting of a group of neurons spanning five early layers and encompassing around 50 thousand parameters, is a compelling example of an ERE that can only be rendered accessible through a pattern recognition process seeking MI. To uncover this circuit, Olah and colleagues utilise various epistemic activities including decomposition and localisation. They begin by using feature visualisation, a technique in which an image is synthesised to maximise the activation of a given ANN component (Olah et al., 2017).[12] The Distill team uses the term *features* to mean human interpretable concepts that components in ANNs become functionally specialised for.[13] They taxonomise the neurons in the first 5 convolutional layers of InceptionV1 into layer-wise *families* (i.e., functional groups) based on which features they are sensitive to (e.g., Gabor filters, colour contrasts, lines, curves). Using feature visualisations to represent neurons and families in diagrams of the ANN makes reasoning about their functional relationship significantly more tractable. Notice that this epistemic activity is composed of iterating multiple epistemic operations: producing visualisations for each neuron, applying labels to them, and grouping the neurons into families.

[11] In biological brains, each neuron's weights have to be learned separately, even if they are learning to detect the same visual feature (e.g., edges). In ANNs, exploiting the assumption that features may appear anywhere in any image, weights can be shared among neurons for computational efficiency.

[12] For reasons of space and to retain focus on the pattern recognition practice, we will not go into formal details but instead paint the general picture in relatively broad strokes. For technical details see the referenced work.

[13] The authors conceive of features as corresponding to *directions in neuronal activation space*. In theory, these directions can be defined across the vector space determined by the activity of arbitrary groups of neurons.

Taxonomising neurons serves as a useful starting point, but a mechanistic explanation of InceptionV1 also requires characterising the *interactions* between neurons. To achieve this, having documented the presence of curve detector neurons in InceptionV1's third layer, the investigators followed their connections to see how upstream neurons (those closer to the input layer) were contributing to their capabilities. This required the application of additional tools, such as methods to visualise the connection weights between neurons (Voss et al., 2021). Iterating this strategy all the way back to the input layer allowed the researchers to develop a holistic view of the circuit. In so doing, Olah and colleagues uncovered further meso-scale EREs such as *circuit motifs*, which are recurring patterns across multiple neurons. For example, many neurons are "rotationally equivariant", detecting features (e.g., curves) which are identical apart from their spatial orientation (Olah et al., 2020b). Ultimately, the Distill team concisely describes the mechanistic structure of the curve circuit as follows: "Gabor filters turn into proto-lines which build lines and early curves. Finally, lines and early curves are composed into curves" (Cammarata et al., 2021). Through coordinated and systematic application of epistemic activities, the full curve circuit becomes recognisable as an ERE distinct from the rest of the network within which it is embedded.

To ensure that discovered patterns are robust, pattern recognition must adhere to epistemic norms such as integrating multiple sources of evidence and ruling out alternative explanations.[14] Otherwise, overreliance on visualisation techniques which involve researcher degrees of freedom risks falling foul of common scientific blunders like confirmation bias (Freiesleben, 2024; Gelman & Loken, 2019; Pu & Kay, 2018). In line with this, the Distill researchers used multiple techniques to verify their hypotheses. For example, they observed the responses of curve detector neurons to natural images to confirm they played their hypothesised roles with real stimuli. However, this raises the possibility of a rival hypothesis; could the curve detectors actually specialise in detecting a finite set of specific curved objects, rather than the more general sub-function of *curve detection*? To exclude this rival hypothesis, Olah and colleagues systematically tested the neurons with synthetic stimuli, determining that the curve detection neurons in layer 3b of InceptionV1 are invariant to both fill and colour (i.e., they really are curve detectors).

In addition to applying decomposition and localisation to the trained network, Olah and colleagues carried out a (computational) recomposition of the curve detection circuits they had investigated (Cammarata et al., 2021). To do this, one of the authors, leveraging the insights gleaned from the discovery process, designed a curve detection algorithm with the same structure as InceptionV1's by hand. That is to say, they manually set the values of the weights between the neurons. Having done so, the behaviour of the hand-designed network could be compared to that of InceptionV1 by exposing them to identical stimuli and observing their responses. A highly similar response profile was demonstrated, suggesting that Olah and colleagues really had recomposed the curve detection circuit. Indeed, we think this

[14] For more on robustness see, e.g., Wimsatt (1981).

recomposition constitutes a sufficiently *severe test* to be confident in Olah and colleagues' hypothesis of the functional organisation of the curve circuit (see Lam, 2022 for a similar claim).[15]

The Distill investigation into InceptionV1 exemplifies all of the properties of a pattern recognition practice, viz., the coordinated application of epistemic operations consisting of a shared set of concepts, tools, and skills.[16] Olah and colleagues organised their investigation around core concepts: *features* and *circuits.*[17] When it comes to tools, various visualisation techniques, synthetic stimuli, and explainability interfaces were developed and employed throughout the discovery process. Finally, implementing a pattern recognition practice on ANNs demands multiple skills. Unlike in a wet lab, manipulation of physical instruments is not required. Rather, a firm grasp of the mathematics underlying ANNs is critical. What *is* shared with discovery processes in the life sciences are the reasoning skills required, i.e., choosing which system properties to track and visualise, abductive reasoning to generate hypotheses, and design of experimental procedures to test those hypotheses.

3.3 Benefits of the MI approach

Unlike the divide-and-conquer strategy, the pattern recognition practice applied by Olah and his colleagues followed a coherent, systematic research agenda. The scientists' goal was to uncover functional components within the classifier and to illuminate how they work together in the system's overall functional architecture. To this end, they (i) *systematically* searched for functional units (i.e., EREs) that had not been explicitly coded, (ii) *coordinated* different epistemic activities to help reveal the relationships between different measurements, and (iii) *iteratively* refined hypotheses concerning various EREs, their activities, and their interactions. Eventually, they stitched together a mechanistic explanation rendering intelligible how the first five layers of InceptionV1 work. With respect to curve detection specifically, the authors note that "although curve detection involves more than 50,000 parameters,

[15] A severe test of a hypothesis H, as characterised by Deborah Mayo, is such that "H agrees with the data (H passes the test), [and] also […] with high probability, H would not have passed the test so well, were H false" (Mayo, 2018, p. 92). We think that the potential for MI to provide severe tests of hypotheses about the internal structure of trained models distinguishes it from the divide-and-conquer strategy (see Geiger et al., 2023 for another approach to MI that arguably implements severe testing).

[16] Note that in some cases the computational methods described above have direct analogues in biology and neuroscience. However, the similarity we wish to evoke holds at the level of *research strategies* not at the level of *individual methods*. For discussion of specific methodological similarities see Ivanova et al. (2021) on how probing methods in MI were derived from multivariate pattern analysis in neuroscience and see, e.g., Bashivan et al. (2019) for use of synthetic visualisations in a neuroscience context. On how interventionist theories of causality can be applied to test causal hypotheses in ANNs see Geiger et al. (2021).

[17] The investigation of *circuits* as a relevant unit of analysis has been adopted by further researchers (Marks et al., 2024; Wang et al., 2022), highlighting the cumulative nature of MI research.

those parameters actually implement a simple algorithm that can be read off the weights and described in just a few English sentences" (Cammarata et al., 2021).

The case of Distill highlights that MI research has two distinct advantages. First, it can shed light on EREs (e.g., the curve detection circuit) of opaque AI systems that are inaccessible through individual XAI methods. There are two reasons for this. The first is that, like measurement tools in the life sciences, individual XAI methods are tuned to picking up specific kinds of information. While tools specialised for answering specific questions are clearly useful, they are also crucially limited. For instance, EEG measures electrical potentials, but it cannot tell us about brain structure. The second reason is that while a lot of information may be gathered from applying different XAI methods, that information will often stand unconnected. By contrast, successful pattern recognition (yielding ERE identification) crucially depends on the coordinated application of many relevant instruments and the *integration* of multiple sources of evidence in conjunction with careful *scientific reasoning* (cf., Kästner, 2018), potentially involving severe testing to rule out alternative hypotheses (see fn 15). By identifying previously unknown EREs, and confirming they have been characterised accurately (e.g., through recomposition), researchers pursuing MI research will often be rewarded with deeper explanations revealing a wider set of EREs and uncovering the functional architecture of AI systems (see Fig. 1). We think this is precisely what is needed to render AI systems holistically explainable.

This leads us to the second advantage of MI research. Characterising a system's functional structure in terms of how relevant components work together enables domain experts to control (e.g., through targeted interventions) and predict system behaviour even for novel situations (Geiger et al., 2023; Meng et al., 2023; Zou et al., 2023). As such, it helps prevent unexpected failures and thus contributes to satisfying important societal desiderata. We already briefly outlined how MI contributes to system reliability and safety in high-stakes domains (in Section 2). While image recognition may not be such a domain per se, the case of Distill serves as proof-of-principle that a fine-grained characterisation of the functional organisation of complex ML-based systems is possible. Besides, ML-based image recognition is likely to be utilised by AI systems employed in high-stakes domains such as traffic control, law, and surveillance. In addition to ensuring safety, the kind of understanding MI elicits is also crucial to support trust. While the relationship between interpretability, trust, and trustworthiness is complicated (Markus et al., 2021), it is widely agreed that if AI systems are going to be employed to make socially and morally consequential decisions, they need to be *trustworthy* (Kästner et al., 2021). The trust that laypeople have in technological systems is often founded upon their belief that *someone, somewhere* understands the system deeply (Sloman & Rabb, 2016). For sufficiently complex AI systems, we believe, such deep understanding depends upon uncovering EREs that are not attainable through the divide-and-conquer strategy. By contrast, MI has the potential to produce the kind of expert understanding that supports trust.

Another desideratum for which MI is likely to be important is scientific understanding. ANNs have already had a large impact on many scientific fields (Boge, 2022; Bouatta et al., 2021; Cichy & Kaiser, 2019). They achieved unrivalled predictive success for complex phenomena as diverse as neural activity, protein

folding, and particle physics. However, philosophers have expressed concern that ANN models, while undoubtedly impressive, do not confer scientific understanding (Chirimuuta, 2021). This limitation is usually traced to the inscrutable learned structure of these models. However, this structure is *precisely* what MI seeks to characterise – and once it is captured successfully, formerly opaque ML-based systems may be re-engineered as transparent or interpretable ones. Therefore, we expect MI to play an increasingly important role for scientific research in the coming years (see also, Crook & Kästner, Forthcoming).

Though our discussion in this section has focused on the potential of MI, we want to be clear that we are *not* calling to eliminate traditional XAI. Rather, we take both approaches to be *complementary*. MI research requires extensive technical knowledge, scientific skill, financial resources, and time. Just as in life science research, it may take several iterations before a relatively stable mechanistic explanation for a phenomenon is uncovered (cf. Craver & Darden, 2013; Kästner & Haueis, 2021). Thus, seeking MI for AI systems will usually incur much greater investment (both financial and labour) than developing XAI methods and applying them in well-defined contexts. These higher costs might only be worth paying in high-stakes scenarios. Likewise, understanding the overall functional organisation of a system might simply not be relevant in some cases (Durán & Jongsma, 2021). Besides, we think that both MI research and applying traditional XAI methods in a divide-and-conquer fashion will often be *mutually supportive*. Researchers seeking MI can employ specific XAI methods as tools to support some of their coordinated epistemic activities. Likewise, MI research may help refine specific XAI methods, e.g. by revealing new EREs, or improve the mapping between different XAI methods and contexts for the divide-and-conquer strategy. Thus, to what extent MI research and traditional XAI research will be required or useful will depend on the questions at hand, the stakeholders involved, and the specific desiderata at play. Still, it is important that MI research and its unique potential is not overlooked – neither by practitioners nor theorists of XAI.

4 Six worries … and responses

Before closing, we shall briefly outline and respond to some possible objections to the MI approach. Though the list below is not exhaustive, we take it to capture the most common worries our proposal will face. We intentionally keep this brief, as a full discussion would require a distinct project.

First, some critics might worry that the chances of successful MI research are too low. For instance, because (#1) the *complexity of large models* (due to their number of parameters and non-linearities) makes explaining them mechanistically hopeless (e.g., Cearns et al., 2019). Consider GPT3, which contains 175 billion parameters (Brown et al., 2020). Clearly, manually investigating every parameter of such large models one-by-one is infeasible. However, this is not required for MI. Instead, researchers are seeking patterns within the complex system to characterise its function. Crucially, these patterns may involve many parameters, and they can also recur multiple times within the system, dramatically reducing effective complexity (recall

the neuron families and circuit motifs discussed in Section 3.2). There is also the potential for some aspects of MI research to be automated, further ameliorating the problem (Conmy et al., 2023; Hernandez et al., 2022). Besides, models need not be explained exhaustively "all the way down" to parameter-level to accrue important benefits. The diversity of societal desiderata for which mechanistic insights may be relevant precludes specifying general criteria for sufficient mechanistic understanding. However, in some cases, relatively abstract descriptions may be sufficient (e.g., Tenney et al., 2019; see also Lindsay & Bau, 2023 on evaluating understanding of neural systems; and Craver & Kaplan, 2020 on sufficiency conditions for mechanistic explanation more generally). This is well in line with the demands formulated in the EU AI Act (see Section 2) as well as considerations about explainability and understanding being relative to different stakeholder needs (Langer et al., 2021a, 2021b).

Others may worry that (#2) we might be *lacking the right concepts* to accurately characterise an AI system's functional organisation, either in terms of the complex features it exploits (Boge, 2022)[18] or in terms of how it represents and processes those features. We think that even if this is true for now, it is not an argument against seeking MI in principle. It can make implementing a pattern recognition practice *harder*, but forming new concepts is part of the ordinary business of scientific discovery (cf. Craver & Darden, 2013, chapter 5). Besides, the entire history of the life sciences speaks to humans' ability to characterise unfamiliar and complex domains in various useful ways (Bechtel & Richardson, 1993; S. D. Mitchell, 2002). Indeed, Olah and colleagues borrow and fruitfully apply the concept of *network motifs* from systems biology during their curve circuits investigation (Alon, 2006; Olah et al., 2020b). Thus, we see little reason (theoretical or empirical) to think that trained AI models contain structure so alien that human beings are *fundamentally* unable to grasp it.[19] Where concepts are lacking, they can be invented (see Schubert et al., 2021 for an example of this in Distill's work). Science frequently operates by making domains interpretable through visualisation, simplification, idealisation, and abstraction (Fleisher, 2022; Levy & Bechtel, 2013). To be sure, the conceptual challenge here should not be taken lightly, but we think the state of evidence suggests that MI is at least worth attempting.

Besides, a critic might protest that (#3) the analogy to life sciences research is flawed because *we do not actually understand how living systems work either*. Indeed, many neuroscientists admit that we do not yet have a comprehensive mechanistic understanding of how the brain works despite decades of effort (Buzsáki, 2020; Pessoa, 2023). Yet, we see reasons to be *more* optimistic about understanding AI systems through MI than explaining the brain: (i) since AI systems are designed

[18] Boge's concern here is with *w-opacity*, that is, opacity with respect to whatever information hidden in the data the trained system has learned to exploit. W-opacity is thus narrower than the functional characterisation targeted by MI, which *also* involves specifying *how* the subparts of a system encode and process that learned information to generate behaviour. In other words: w-opacity is part of the problem MI aims to overcome.

[19] We acknowledge that this is a complicated issue and much turns on how one construes "grasp" in this sense. However, discussing this in detail would make for another paper. For discussions on the concept of grasping see, e.g., Baumberger et al. (2016) and Janvid (2018).

and trained by us, we have exhaustive knowledge about their architectures, learning algorithms, objective functions, training data, and learned connection weights, (ii) we can perform any conceivable experiment with arbitrary precision and complete access to the causal consequences of our interventions, and (iii) we can choose to develop AI systems which are tailored to be amenable to scientific investigation (e.g., by enforcing linearity wherever it does not significantly hurt model performance). While none of these advantages makes applying MI to complex AI systems *easy*, we think that in combination they warrant cautious optimism.

Similarly, critics may worry that (#4) the analogy to life science research is flawed because *ANNs defy decomposition into components with different functional roles*. We grant that, given the distributed nature of ANNs, one should not assume subfunctions will map neatly onto pre-individuated structures within an ANN (neurons, layers, etc.). However, as discussed in Section 3.1, this is not a pre-requisite for MI to prove valuable. Identifying components relevant for a specific function may involve characterising more exotic and distributed structures (Bricken et al., 2023). This may not be straightforward, but that is another worry (akin to #1 and #2). The exact criteria a system must satisfy to afford effective mechanistic discovery processes is a complex topic we cannot address in this paper (for classic and recent discussions see Simon, 2008; Woodward, 2013; Zednik, 2015). Broadly speaking though, we think extant evidence (including the Distill case study) strongly suggests that ANNs and other modern AI systems tend to be decomposable *enough* for MI to provide value. Further, as Zednik (2015) points out, the ongoing production of more powerful methods and strategies for conducting mechanistic discovery processes continually expands the range of systems that are usefully decomposable. Overall, we consider decomposability a matter of degree (cf. Bechtel & Richardson, 1993). And precisely how (and how easily) decomposable specific ANNs turn out to be is an empirical question (that only seriously attempting MI can possibly answer).

Even if MI research is successful, some might worry that the benefits we highlighted above (see Sections 2 and 3.3) will not be worth the trouble. For instance, those sympathetic to our approach in principle may argue that (#5) *it will be impossible for researchers to seek MI for large AI systems as quickly as they are developed*. It would just be too resource intensive. While we acknowledge MI research requires significant resources (see Section 3.3), we think there are good reasons to pursue it anyway. First, it might simply be worth the investment if otherwise we cannot safely deploy AI in high-stakes scenarios (see Section 2). Second, because mechanistic insights may be transferable from one system to another, the investment might pay off more quickly than we think. For instance, insights about the functional organisation of one language model may well transfer to other language models (*mutatis mutandis* for other domains). Theoretical reasons for this optimism are provided by Cao and Yamins' (2021) *contravariance principle*, which implies that functional convergence should be expected in systems that can perform complex tasks efficiently (see also Bansal et al., 2021; Schrimpf et al., 2020). Besides, researchers are increasingly converging on architectural choices for AI systems (Bommasani et al., 2022). Since model architecture constrains both the kind of functional structure that can emerge and the kind of epistemic operations researchers can apply, this is positive news for MI. Moreover, the resource intensive nature of

pattern recognition practices can be viewed as a feature, rather than a bug. There is still understandable resistance to large AI systems being deployed in high-stakes scenarios (Rudin, 2019). In light of these concerns, it may be reasonable to propose some degree of mechanistic explainability as a pre-requisite for adopting technological innovations that could prove consequential and difficult to reverse (Stirling, 2007). A time-consuming requirement could give slow political processes time to react and inject democratic influence into decisions that impact all of society (Bender et al., 2021; Floridi et al., 2018).

Other critics may worry that (#6) *MI will not yield accurate predictions of AI systems' behaviours in social contexts* – though this is what really matters when we think about deploying AI in modern society (Bender et al., 2021). We acknowledge MI will not be sufficient to precisely predict all effects of deploying models in specific real-world contexts with idiosyncratic data distributions (Quiñonero-Candela, 2009; such prediction is generally intractable, see Shalizi, 2006). However, given MI research seeks to reveal how AI systems work as a whole, it yields deeper and more generalisable insights into how a system may behave in novel (social) contexts than applying the divide-and-conquer strategy. Further, the ability to intervene on trained systems, which MI purports to deliver, may prove important for controlling AI systems after they are deployed. Thus, even if MI is costly and the benefits are not all-encompassing, it may still be superior to its alternatives.

Before concluding, we wish to stress that our aim in this paper is to characterise MI as a distinctive approach to XAI, not to present it as a panacea. As such, though we reject the notion that any of these objections is fatal to the MI research program, we acknowledge that they are legitimate considerations which must be factored into a comprehensive assessment of its pursuit worthiness in specific cases. In a nutshell, MI research can offer answers to research questions which are in principle unavailable to those pursuing a divide-and-conquer strategy (see Section 2). As such, it is important for practitioners to pursue MI to gain explainability where other strategies are unproductive. And it is also crucial for theorists to accommodate MI in their portrayals of the XAI research landscape (see Section 3.3). For only by having the full range of strategies for explaining AI systems in view can we make appropriate choices about which research to pursue to ensure societal desiderata are satisfied.

5 Conclusion

Contemporary XAI research is seeking explainability for opaque AI systems in a divide-and-conquer fashion, viz. by employing specific XAI methods in various contexts. While this strategy has its merits, it fails to illuminate how trained AI systems work as a whole. Yet precisely this kind of holistic understanding is needed to satisfy important desiderata placed on AI systems by society. We argued that mechanistic interpretability research, though it is resource intensive, can serve as a remedy: by uncovering a wider range of EREs than any individual XAI method, MI contributes to a deeper and more holistic understanding of complex AI systems' functioning. It aims to uncover the functional organisation of trained AI systems, including both the complex features they exploit and how they represent and process

those features. These insights can be utilised to predict and control system behaviour beyond pre-defined contexts, which is crucial to satisfy societal desiderata. Appreciating these unique benefits of MI research will help scholars to make adequate choices about what research to pursue to ensure we can safely and reliably deploy ML-based systems in modern society.

Acknowledgements We are indebted to the members of the project "Explainable Intelligent Systems (EIS)", the participants of the Philosophy Breakfast in Bayreuth, and the audiences at GAP.11 and the Cambridge Philosophy of Science Lectures for feedback on earlier versions of this paper. For detailed comments we are especially thankful to Marta Halina, Konstantinos Voudouris, Johanna Thoma, Marius Backmann, Olivier Roy, Paolo Galeazzi, Sara Mann, Andreas Sesing-Wagenpfeil, Eva Schmidt, Florian Boge, and two anonymous reviewers.

Author contribution Original research paper; authors contributed equally.

Funding Open Access funding enabled and organized by Projekt DEAL. Volkswagen Foundation Grant AZ 9B830.

Data Availability N/A.

Declarations

Ethical approval N/A.

Informed consent N/A.

Conflict of interest None.

References

Alon, U. (2006). An introduction to systems biology: Design principles of biological circuits. *Chapman and Hall/CRC*. https://doi.org/10.1201/9781420011432

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety* (arXiv:1606.06565). arXiv. https://doi.org/10.48550/arXiv.1606.06565

Artificial Intelligence Act. (2024). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf. Accessed 23 Sept 2024.

Baetu, T. M. (2011). Mechanism schemas and the relationship between biological theories. In P. McKay Illari, F. Russo, & J. Williamson (Eds.), *Causality in the Sciences* (pp. 407–424). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199574131.003.0019

Bansal, Y., Nakkiran, P., & Barak, B. (2021). Revisiting model stitching to compare neural representations. *Advances in Neural Information Processing Systems*, *34*, 225–236.

Baraniuk, R., Donoho, D., & Gavish, M. (2020). The science of deep learning. *Proceedings of the National Academy of Sciences, 117*(48), 30029–30032. https://doi.org/10.1073/pnas.2020596117

BarredoArrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science (New York, N.Y.)*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). *Network dissection: Quantifying interpretability of deep visual representations*. 3319–3327. https://doi.org/10.1109/CVPR.2017.354

Baumberger, C., Beisbart, C., & Brun, G. (2016). What is understanding? An overview of recent debates in epistemology and philosophy of science. In *Explaining Understanding*. Routledge.

Bechtel, W. (2009). Looking down, around, and up: mechanistic explanation in psychology. *Philosophical Psychology*, *22*. https://doi.org/10.1080/09515080903238948

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., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research* (pp. xiv, 286). Princeton University Press.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data*, *4*.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Boge, F. J. (2022). Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines, 32*(1), 43–75. https://doi.org/10.1007/s11023-021-09569-4

Boge, F. J. (2023). Functional Concept Proxies and the Actually Smart Hans Problem: What's Special About Deep Neural Networks in Science. *Synthese, 203*(1), 16. https://doi.org/10.1007/s11229-023-04440-8

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models*. https://doi.org/10.48550/arXiv.2108.07258

Bouatta, N., Sorger, P., & AlQuraishi, M. (2021). Protein structure prediction by AlphaFold2: Are attention and symmetries all you need? *Acta Crystallographica. Section d, Structural Biology, 77*(Pt 8), 982–991. https://doi.org/10.1107/S2059798321007531

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., … Olah, C. (2023). Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Brigandt, I. (2011). Explanation in biology: Reduction, pluralism, and explanatory aims. *Science & Education, 22*(1), 69–91. https://doi.org/10.1007/s11191-011-9350-7

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Burnston, D. C. (2021). Getting over atomism: Functional decomposition in complex neural systems. *The British Journal for the Philosophy of Science, 72*(3), 743–772. https://doi.org/10.1093/bjps/axz039

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Buzsáki, G. (2020). The brain–cognitive behavior problem: A retrospective. *eNeuro*, *7*(4). https://doi.org/10.1523/ENEURO.0069-20.2020

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., & Schubert, L. (2020). Thread: Circuits. *Distill*, *5*(3). https://doi.org/10.23915/distill.00024.

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

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 2 -- constraint-based intelligibility. https://arxiv.org/pdf/2104.01489. Accessed 23 Sept 2024.

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*(1), Article 1. https://doi.org/10.1038/s41398-019-0607-2

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese, 199*(1), 767–790. https://doi.org/10.1007/s11229-020-02713-0

Chollet, F. (2021). *Deep Learning with Python* (2nd ed.). Simon and Schuster.

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences, 23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

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

Craver, C. F. (2001). Role functions, mechanisms, and hierarchy. *Philosophy of Science, 68*(1), 53–74.

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

Craver, C. F., & Darden, L. (2013). *In Search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science, 71*(1), 287–319. https://doi.org/10.1093/bjps/axy015

Crook, B., & Kästner, L. (Forthcoming). Don't Fear the Bogeyman: On Why There is No Prediction-Understanding Trade-Off for Deep Learning in Neuroscience. In J. M. Durán & G. Pozzi (Eds.), *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Springer.

Dennett, D. C. (1991). Real patterns. *Journal of Philosophy, 88*(1), 27–51. https://doi.org/10.2307/2027085

Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, medethics-2020–106820. https://doi.org/10.1136/medethics-2020-106820

Eden, A. H. (2007). Three paradigms of computer science. *Minds and Machines, 17*(2), 135–167. https://doi.org/10.1007/s11023-007-9060-8

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition. In *arXiv e-prints*. https://doi.org/10.48550/arXiv.2209.10652

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Fawzi, A., Fawzi, H., & Fawzi, O. (2018). Adversarial vulnerability for any classifier. *Advances in Neural Information Processing Systems*, *31*.

Fleisher, W. (2022). Understanding, idealization, and explainable AI. *Episteme, 19*(4), 534–560. https://doi.org/10.1017/epi.2022.39

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Freiesleben, T. (2024). *Artificial neural nets and the representation of human concepts*. https://doi.org/10.48550/arXiv.2312.05337

Geiger, A., Lu, H., Icard, T., & Potts, C. (2021). Causal abstractions of neural networks. *Advances in Neural Information Processing Systems, 34*, 9574–9586.

Geiger, A., Potts, C., & Icard, T. (2023). *Causal abstraction for faithful model interpretation*. https://doi.org/10.48550/arXiv.2301.04709

Geiger, A., Wu, Z., Lu, H., Rozner, J., Kreiss, E., Icard, T., Goodman, N., & Potts, C. (2022). Inducing causal structure for interpretable neural networks. *Proceedings of the 39th International Conference on Machine Learning*, 7324–7338.

Gelman, A., & Loken, E. (2019). *The garden of forking paths: Why multiple comparisons can be a problem , even when there is no " fishing expedition " or " p-hacking " and the research hypothesis*

*was posited ahead of time **. http://www.stat.columbia.edu/~gelman/research/unpublished/forking.pdf. Accessed 23 Sept 2024

Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal neurons in artificial neural networks. *Distill*, *6*(3). https://doi.org/10.23915/distill.00030.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 93:1–93:42. https://doi.org/10.1145/3236009

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. https://doi.org/10.1145/3593013.3594067

Haugeland, J. (1998). *Having Thought: Essays in the Metaphysics of Mind*. Harvard University Press.

Hendrycks, D., Mazeika, M., & Dietterich, T. G. (2018). Deep anomaly detection with outlier exposure. *ArXiv*. https://www.semanticscholar.org/reader/2d8c97db4bae00ff243d122b957091a236a697a7. Accessed 23 Sept 2024.

Hendrycks, D., Mazeika, M., & Woodside, T. (2023). *An Overview of Catastrophic AI Risks*. https://doi.org/10.48550/arXiv.2306.12001

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., & Andreas, J. (2022). *Natural Language Descriptions of Deep Visual Features*. https://doi.org/10.48550/arXiv.2201.11114

Howick, J., Glasziou, P., & Aronson, J. K. (2010). Evidence-based mechanistic reasoning. *Journal of the Royal Society of Medicine, 103*(11), 433–441. https://doi.org/10.1258/jrsm.2010.100146

Huang, S., Yang, J., Fong, S., & Zhao, Q. (2020). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters, 471*, 61–71. https://doi.org/10.1016/j.canlet.2019.12.007

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

Ivanova, A. A., Hewitt, J., & Zaslavsky, N. (2021). *Probing artificial neural networks: Insights from neuroscience*. https://doi.org/10.48550/arXiv.2104.08197

Janvid, M. (2018). Getting a grasp of the grasping involved in understanding. *Acta Analytica, 33*(3), 371–383. https://doi.org/10.1007/s12136-018-0348-5

Jugovac, M., & Jannach, D. (2017). Interacting with Recommenders—Overview and Research Directions. *ACM Transactions on Interactive Intelligent Systems*, *7*(3), 10:1–10:46. https://doi.org/10.1145/3001837

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2022). Towards Causal Algorithmic Recourse. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, & W. Samek (Eds.), *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 139–166). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_8

Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A, 68*, 68–79. https://doi.org/10.1016/j.shpsa.2018.01.011

Kästner, L., & Haueis, P. (2021). Discovering patterns: On the norms of mechanistic inquiry. *Erkenntnis, 86*(6), 1635–1660. https://doi.org/10.1007/s10670-019-00174-7

Kästner, L., Langer, M., Lazar, V., Schomacker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 169–175. https://doi.org/10.1109/REW53955.2021.00031

Keil, F. (2019). *How do partial understandings work?* (pp. 191–208). https://doi.org/10.1093/oso/9780190860974.003.0010

Keogh, E., & Mueen, A. (2017). Curse of Dimensionality. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning and Data Mining* (pp. 314–315). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_192

Kirsch, A. (2017). *Explain to whom? Putting the User in the Center of Explainable AI*. CEx@AI*IA. https://ceur-ws.org/Vol-2071/CExAIIA_2017_keynote_1.pdf. Accessed 23 Sept 2024.

Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a non-functional requirement. *2019 IEEE 27th International Requirements Engineering Conference (RE)*, 363–368. https://doi.org/10.1109/RE.2019.00046

Lam, N. (2022). Explanations in AI as claims of tacit knowledge. *Minds and Machines, 32*(1), 135–158. https://doi.org/10.1007/s11023-021-09588-1

Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., & Speith, T. (2021a). Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment, 29*(2), 154–169. https://doi.org/10.1111/ijsa.12325

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021b). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-08987-4

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of Science, 80*(2), 241–261.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2023). *Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task*. https://doi.org/10.48550/arXiv.2210.13382

Lindsay, G. W., & Bau, D. (2023). Testing methods of neural systems understanding. *Cognitive Systems Research, 82*, 101156. https://doi.org/10.1016/j.cogsys.2023.101156

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences, 117*(48), 30046–30054. https://doi.org/10.1073/pnas.1907367117

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2024). *Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models*. https://doi.org/10.48550/arXiv.2403.19647

Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics, 113*, 103655. https://doi.org/10.1016/j.jbi.2020.103655

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781107286184

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2023). *Locating and Editing Factual Associations in GPT*. https://doi.org/10.48550/arXiv.2202.05262

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). *Ablation Studies in Artificial Neural Networks*. https://doi.org/10.48550/arXiv.1901.08644

Millière, R., & Buckner, C. (2024). *A Philosophical Introduction to Language Models -- Part I: Continuity With Classic Debates*. https://doi.org/10.48550/arXiv.2401.03910

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

Mitchell, S. D. (2002). Integrative Pluralism. *Biology and Philosophy, 17*(1), 55–70. https://doi.org/10.1023/A:1012990030867

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book/. Accessed 23 Sept 2024.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). *Progress measures for grokking via mechanistic interpretability*. https://doi.org/10.48550/arXiv.2301.05217

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126. https://doi.org/10.1145/360018.360022

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction, 27*(3–5), 393–444. https://doi.org/10.1007/s11257-017-9195-0

Nyrup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology, 24*(1), 13. https://doi.org/10.1007/s10676-022-09632-3

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, *5*(3), e00024.001. https://doi.org/10.23915/distill.00024.001

Olah, C., Cammarata, N., Voss, C., Schubert, L., & Goh, G. (2020). Naturally Occurring Equivariance in Neural Networks. *Distill*, *5*(12), e00024.004. https://doi.org/10.23915/distill.00024.004

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, *2*(11). https://doi.org/10.23915/distill.00007

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*, *3*(3). https://doi.org/10.23915/distill.00010

Páez, A. (2019). The pragmatic turn in Explainable Artificial Intelligence (XAI). *Minds and Machines, 29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Pavlick, E. (2023). Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society a: Mathematical, Physical and Engineering Sciences, 381*(2251), 20220041. https://doi.org/10.1098/rsta.2022.0041

Pessoa, L. (2023). The entangled brain. *Journal of Cognitive Neuroscience, 35*(3), 349–360. https://doi.org/10.1162/jocn_a_01908

Pu, X., & Kay, M. (2018). The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics : Position Paper. *IEEE Evaluation and beyond - Methodological Approaches for Visualization (BELIV), 2018*, 37–45. https://doi.org/10.1109/BELIV.2018.8634103

Quiñonero-Candela, J. (Ed.). (2009). *Dataset shift in machine learning*. MIT Press.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., & Sohl-Dickstein, J. (2017). *On the expressive power of deep neural net-works*. https://arxiv.org/pdf/1606.05336.pdf. Accessed 23 Sept 2024.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., & Nanda, N. (2024). *Improving Dictionary Learning with Gated Sparse Autoencoders*. https://doi.org/10.48550/arXiv.2404.16014

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks*. https://doi.org/10.48550/arXiv.2207.13243

Räz, T. (2023). Methods for identifying emergent concepts in deep neural networks. *Patterns, 4*(6), 100761. https://doi.org/10.1016/j.patter.2023.100761

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 97–101. https://doi.org/10.18653/v1/N16-3020

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), Article 1. https://doi.org/10.1609/aaai.v32i1.11491

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach (4th Edition)*. Pearson.

Schrimpf, M., Kubilius, J., Lee, M. J., RatanMurty, N. A., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron, 108*(3), 413–423. https://doi.org/10.1016/j.neuron.2020.07.040

Schubert, L., Voss, C., Cammarata, N., Goh, G., & Olah, C. (2021). High-low frequency detectors. *Distill*, *6*(1), e00024.005. https://doi.org/10.23915/distill.00024.005

Shalizi, C. R. (2006). Methods and Techniques of Complex Systems Science: An Overview. In T. S. Deisboeck & J. Y. Kresh (Eds.), *Complex Systems Science in Biomedicine* (pp. 33–114). Springer US. https://doi.org/10.1007/978-0-387-33532-2_2

Simon, H. A. (2008). *The sciences of the artificial* (3. ed., [reprint]). MIT Press.

Sloman, S. A., & Rabb, N. (2016). Your understanding is my understanding: Evidence for a community of knowledge. *Psychological Science, 27*(11), 1451–1460. https://doi.org/10.1177/0956797616662271

Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. https://doi.org/10.1145/3351095.3372870

Speith, T. (2022). A review of taxonomies of Explainable Artificial Intelligence (XAI) Methods. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250. https://doi.org/10.1145/3531146.3534639

Steimers, A., & Schneider, M. (2022). Sources of risk of AI systems. *International Journal of Environmental Research and Public Health, 19*(6), 3641. https://doi.org/10.3390/ijerph19063641

Stirling, A. (2007). Risk, precaution and science: Towards a more constructive policy debate. Talking point on the precautionary principle. *EMBO Reports*, *8*(4), 309–315. https://doi.org/10.1038/sj.embor.7400953

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A, 44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601. https://doi.org/10.18653/v1/P19-1452

Voss, C., Cammarata, N., Goh, G., Petrov, M., Schubert, L., Egan, B., Lim, S. K., & Olah, C. (2021). Visualizing Weights. *Distill*, *6*(2), e00024.007. https://doi.org/10.23915/distill.00024.007

Vredenburgh, K. (2022). The right to explanation*. *Journal of Political Philosophy, 30*(2), 209–229. https://doi.org/10.1111/jopp.12262

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3063289

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). *Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small*. https://doi.org/10.48550/arXiv.2211.00593

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). *Jailbroken: How Does LLM Safety Training Fail?*. https://doi.org/10.48550/arXiv.2307.02483

Wimsatt, W. C. (1981). Robustness, Reliability, and Overdetermination (1981). In L. Soler, E. Trizio, T. Nickles, & W. Wimsatt (Eds.), *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science* (pp. 61–87). Springer Netherlands. https://doi.org/10.1007/978-94-007-2759-5_2

Wimsatt, W. C. (1994). The ontology of complex systems: levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy Supplementary, 20*, 207–274. https://doi.org/10.1080/00455091.1994.10717400

Winning, J., & Bechtel, W. (2018). Rethinking causality in biological and neural mechanisms: Constraints and control. *Minds and Machines, 28*(2), 287–310. https://doi.org/10.1007/s11023-018-9458-5

Woodward, J. (2013). Mechanistic explanation: Its scope and limits. *Proceedings of the Aristotelian Society, Supplementary Volumes, 87*, 39–65.

Woodward, J. (2017). Explanation in Neurobiology: An Interventionist Perspective. In D. M. Kaplan (Ed.), *Explanation and Integration in Mind and Brain Science* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780199685509.003.0004

Zednik, C. (2015). Heuristics, Descriptions, and the Scope of Mechanistic Explanation. In P.-A. Braillard & C. Malaterre (Eds.), *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences* (pp. 295–318). Springer Netherlands. https://doi.org/10.1007/978-94-017-9822-8_13

Zednik, C. (2021). 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

Zhou, J., Gandomi, A. H., Chen, F., & Holzinger, A. (2021). Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, *10*(5), Article 5. https://doi.org/10.3390/electronics10050593

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., … Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. https://doi.org/10.48550/arXiv.2310.01405